Neuroscience

Foveated metamers of the early visual system

William F. Broderick ✉, **Gizem Rufo, Jonathan Winawer, Eero P. Simoncelli**

Flatiron Institute, Simons Foundation • Meta, Inc. • Department of Psychology, New York University • Center for Neural Science, New York University • Courant Inst. for Mathematical Sciences, New York University

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
November 1, 2023 (this version)

Posted to bioRxiv
August 2, 2023

Sent for peer review
August 2, 2023

Abstract

Human ability to discriminate and identify visual attributes varies across the visual field, and is generally worse in the periphery than in the fovea. This decline in performance is revealed in many kinds of tasks, from detection to recognition. A parsimonious hypothesis is that the representation of any visual feature is blurred (spatially averaged) by an amount that differs for each feature, but that in all cases increases with eccentricity. Here, we examine models for two such features: local luminance and spectral energy. Each model averages the corresponding feature in pooling windows whose diameters scale linearly with eccentricity. We performed psychophysical experiments with synthetic stimuli to determine the window scaling for which human and model discrimination abilities match, called the critical scaling. We used much larger stimuli than those of previous studies, subtending 53.6 by 42.2 degrees of visual angle. We found the critical scaling for the luminance model was approximately one-fourth that of the energy model, and consistent with earlier studies, that a smaller critical scaling value was required when discriminating a synthesized image from a natural image than when discriminating two synthesized images. We offer a coherent explanation for these results in terms of alignments and misalignments of the models with human perceptual representations.

eLife assessment

This study provides **valuable** insights into how researchers can use perceptual metamers to formally explore the limits of visual representations at different processing stages. While the study is overall **convincing** in terms of approach and results, issues were identified with respect to novelty, sample size, **incomplete** psychophysical methodology, and better motivation of the models tested.

Introduction

Vision science is often concerned with what things look like (appearance), but a long and fruitful thread of research has investigated what humans cannot see, that is, the information they are insensitive to. Perceptual metamers — images that are physically distinct but perceptually indistinguishable — provide direct evidence of such information loss in visual representations. ***Cohen and Kappauf (1985)*** identify this concept in the writings of Isaac Newton, who noted that particular colors of light could be created by mixing other colors together, and trace the word

"metamer" to a 1919 chapter by Wilhelm Ostwald, the first Nobel laureate in chemistry. Color metamers were instrumental in the development of the Young-Helmholtz theory of trichromacy (***Helmholtz, 1852)***. In this context, metamers clarified human sensitivity to light wavelengths, demonstrating that the human visual system projects the infinite dimensionality of the physical signal to three dimensions. It took more than a century before the physiological basis for this — the three classes of cone photoreceptor — was revealed experimentally (***Schnapf et al., 1987)***.

The visual system also discards a great deal of spatial detail, more so in portions of the visual field farthest from the center of gaze. Specifically, the reduction of visual capabilities with increasing eccentricity has been demonstrated for both acuity (***Frisen and Glansholm, 1975)*** and contrast sensitivity (***Banks et al., 1987***; ***Robson and Graham, 1981***; ***Rovamo et al., 1978)***, and is reflected in the physiology: fewer cortical resources are dedicated to the periphery (***Schwartz, 1977)*** and receptive fields in all stages of the visual hierarchy grow with eccentricity (e.g., ***Daniel and Whitteridge (1961)***; ***Dacey and Petersen (1992)***; ***Gattass et al. (1981***, ***1988)***; ***Maunsell and Newsome (1987)***; ***Wandell and Winawer (2015)***). This decrease in acuity has been demonstrated by either scaling the size of features, as in the Anstis eye chart (***Anstis, 1974)***, or by progressively blurring the image, as in ***Anstis (1998)***; ***Thibos (2020)***. More generally, one can explain this decreasing sensitivity to spatial information with "pooling models", which compute local averages of image features in windows that grow larger with eccentricity (***Balas et al., 2009***; ***Freeman and Simoncelli, 2011***; ***Keshvari and Rosenholtz, 2016)***. These models assume that peripheral representations are qualitatively similar to those of the fovea: the same local computations are performed over larger regions. Here, we test two such models — one that averages local luminance (luminance model) and one that averages both local spectral energy and luminance (energy model). We generate images pairs in which one or both images have been manipulated such that the two are model metamers (images that are physically distinct but with identical model representations). The pair of model metamers are also perceptual metamers if the human visual system is insensitive to the differences between them, as schematized in **figure 1** (see ***Watson et al. (1986)*** for an analogous presentation with respect to sensitivity to spatial and temporal frequency). By comparing model and human perceptual metamers, we investigate how well the models' sensitivities (and insensitivities) align with those of the human visual system.

This procedure rests on the assumption that the visual system processes information hierarchically, in a sequence of stages, and information discarded in early stages cannot be recovered by later stages (the "data-processing inequality")[1]. For example, metameric color stimuli produce identical cone responses and thus they cannot be distinguished by any additional downstream neural processing. Perceptually, if two images generate identical responses in all neurons at any stage of processing preceding that of perceptual decision making (e.g., retinal ganglion cells or primary visual cortical neurons), they will appear identical. This is schematized in the central panel of **figure 1**: two images are perceptual metamers if they are indistinguishable to an early visual area N_1 or if this early stage is sensitive to their differences, but those differences are discarded by later stages in the hierarchy. A number of authors have created perceptual metamers by matching complex statistics thought to be represented at a high level of the visual processing hierarchy (***Freeman and Simoncelli, 2011***; ***Keshvari and Rosenholtz, 2016***; ***Wallis et al., 2019***; ***Jagadeesh and Gardner, 2022***; ***Feather et al., 2019)*** and at the level of photoreceptors. However, there are fewer examples of perceptual metamers that match simpler image statistics.

Here, we synthesize metamers for two different models and measure their perceptual discriminability. The two models are based on local pooling of luminance and spectral energy, respectively, which can be loosely associated with two stages of visual physiology. Specifically, retinal ganglion cells encode local light level, and V1 cells encode local spectral energy. For each set of natural images, we used a stochastic gradient descent method to generate model metamers for both models, and measured discrimination capabilities of human observers when comparing the metamers with their corresponding original images, as well as with each other. We also

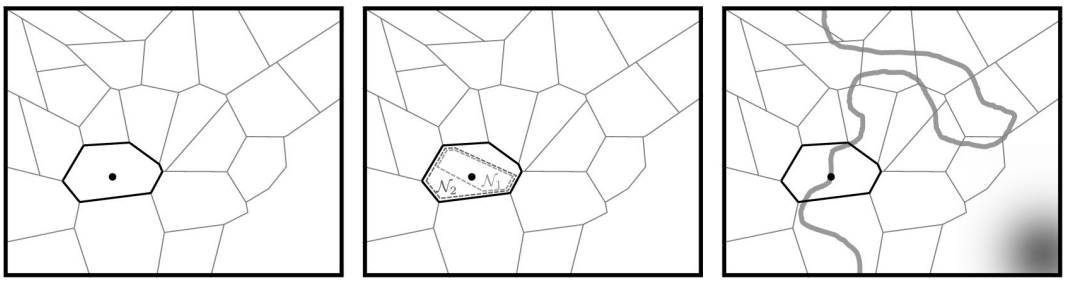


Figure 1.

Schematic diagram of perceptual metamers. Each panel shows a two-dimensional depiction of the set of all possible images: every image corresponds to a point in this space and every point in this space represents an image. Perceptual metamers — images that cannot be reliably distinguished — are equivalence classes in this space and we illustrate this by partitioning the space into distinct perceptually identical regions. Left: example image (black point), and surrounding metameric images (region enclosed by black polygon). Center: In a hierarchical visual system, in which each stage transforms the signals of the previous stage and discards some information, metamers of earlier stages are nested within metamers of later stages. That is, every pair of images that are metamers for an early visual area .N_1 are also metamers for the later visual area N_2. The converse does not hold: there are images that N_1 can distinguish but that N_2 cannot. Right: Two particular image families: Samples of white noise (distribution represented as a grayscale intensity map in the lower right corner) and the set (manifold) of natural images (distribution represented by the curved gray line). Typical white noise samples fall within a single perceptual metamer class (humans are unable to distinguish them). Natural images, on the other hand, are generally distinguishable from each other, as well as from un-natural images (those that lie off the manifold).

examined the influence of the initial image used for the metamer synthesis algorithm. The two types of models and multiple types of comparisons shed light on and raise new questions about what makes images distinguishable.

Results

Foveated pooling models

We constructed foveated models of human perception that capture sensitivity to local luminance and spectral energy (see **figure 3**). Both models are "pooling models" (***Balas et al., 2009***; ***Freeman and Simoncelli, 2011***; ***Keshvari and Rosenholtz, 2016***; ***Wallis et al., 2019)***, which compute statistics as weighted averages over local windows. A specific pooling model is characterized by both the statistics that are pooled and the shapes/sizes of the pooling regions. In the human visual system, receptive field sizes grow proportionally with distance from the fovea, as has been documented in monkey physiology and human fMRI (e.g., ***Gattass et al. (1981)***; ***Wandell and Winawer (2015)***). We reduce this to a single scaling parameter, by assuming smooth overlapping pooling regions that are separable and of constant size when expressed in polar angle and log-eccentricity ("log-polar", which corresponds to the approximate log-polar geometry of visual cortical maps (***Schwartz, 1977)***). The value of this parameter, along with the choice of statistics, determine the sets of model metamers; for a given set of statistics, increasing the scaling value will increase the size of the sets of metameric images, in a nested manner (**figure 2**).

We implemented log-polar pooling windows, with size proportional to their distance from the fovea. Like previous studies (***Freeman and Simoncelli, 2011***; ***Wallis et al., 2019)***, these pooling windows are overlapping and radially-elongated (the radial extent is roughly twice the angular extent), but unlike previous studies, we use Gaussian profiles with more extensive overlap, yielding a smoother representation and higher-quality synthesized images. The proportional overlap between adjacent windows is chosen to alleviate ringing and blocking artifacts in the synthesis, and is held fixed at all scaling values.

In the current study, we examine two models that compute different statistics within their pooling windows. The *luminance* model pools pixel intensities, and thus, a pair of luminance model metamers have the same average luminance within each pooling windows. Since the responses are insensitive to the highest frequencies, luminance model metamers include blurred versions of the target image (in which high frequencies are discarded), but also variants of the target image in which high frequencies are randomized or even amplified. In general, synthesized luminance model metamers will inherit the high frequency information of their initialization image, as can be seen in **figure 4**. The middle row of **figure 4** shows model metamers computed with two example scaling values. While the high-scaling model metamer is clearly perceptually distinct from the target image (regardless of observer fixation location), the low-scaling image is not easily discriminated from the target when fixating at the center of the image (i.e., when the human fovea is aligned with the model fovea).

The *spectral energy* model pools the squared outputs of oriented bandpass filter responses at multiple scales and orientations. It also pools the pixel intensities. The energies are computed using a complex steerable pyramid, which decomposes images into frequency channels selective for 6 different scales and 4 different orientations. Energy is computed by squaring and summing across the real and imaginary responses (arising from even- and odd-symmetric filters) within each channel. These energies, along with the luminances, are then averaged within the spatial pooling windows. Thus, a pair of energy model metamers have the same average oriented energy and luminance within each of these windows. The bottom row of **figure 4** shows model metamers for two different scaling values. The low scaling value for the energy model is approximately matched to the higher scaling value for the luminance model, while the higher

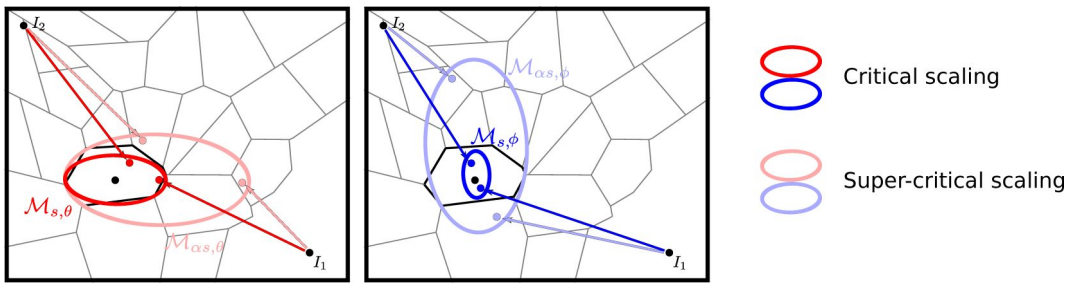


Figure 2.

Left: Pooling models M are parameterized by their scaling value, *s*, and the statistics they pool, *0*. Like any system that discards information, including the human visual system, these models have metamers: sets of images that are perceived as identical, represented graphically as enclosed non-overlapping regions. We draw samples from these sets using an optimization procedure: starting with initial image *Ii* we adjust the pixel values until their pooled statistics match those of the target (original) image *T*. The synthetic image depends on the target image, the metamer model, and also on the initial point and the stochastic synthesis algorithm. For a given set of statistics *0*, increasing the scaling value will increase the size of the metamer set in a nested manner: any image that is a metamer for $M_{s,0}$ will also be a metamer for $M_{\alpha s,0}$, for factor $\alpha > 1$. Right: Changing the set of pooled statistics from *0* to ϕ will result in different sets of model metamers, which may or may not overlap with those of the original model (though both must include the target image, *T*). If the model's metamer classes differ substantially in their shape from the perceptual metamer classes, they will not provide a good description of the perceptual metamers at critical scaling (e.g., the blue ellipse contains only a small portion of the surrounding perceptual metamer region).

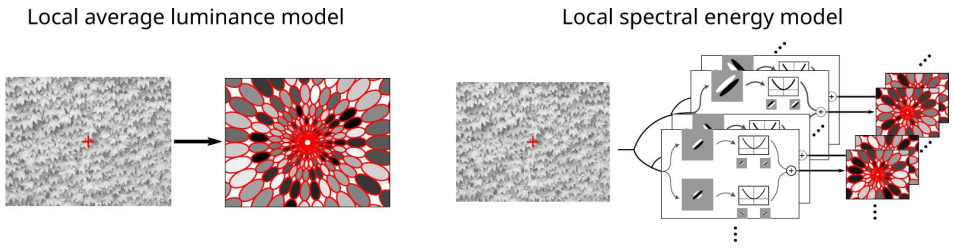


Figure 3.

Two pooling models. Both models compute local image statistics weighted by a Gaussian that is separable in a log-polar coordinate system, such that radial extent is approximately twice the angular extent (half-maximum levels indicated by red contours). Windows are laid out in a log-polar grid, with peaks separated by one standard deviation. A single scaling factor governs the size of all pooling windows. The luminance model (top) computes average luminance, approximating the spatial pooling performed by retinal ganglion cells. The spectral energy model (bottom) computes average spectral energy at 4 orientation and 6 scales, as well as luminance, for a total of 25 statistics per window, approximating the representation of complex cells in primary visual cortex (V1). Spectral energy is computed using the complex steerable pyramid constructed in the Fourier domain (***Simoncelli and Freeman, 1995***), squaring and summing across the real and imaginary components. Full resolution version of this figure can found on the OSF.

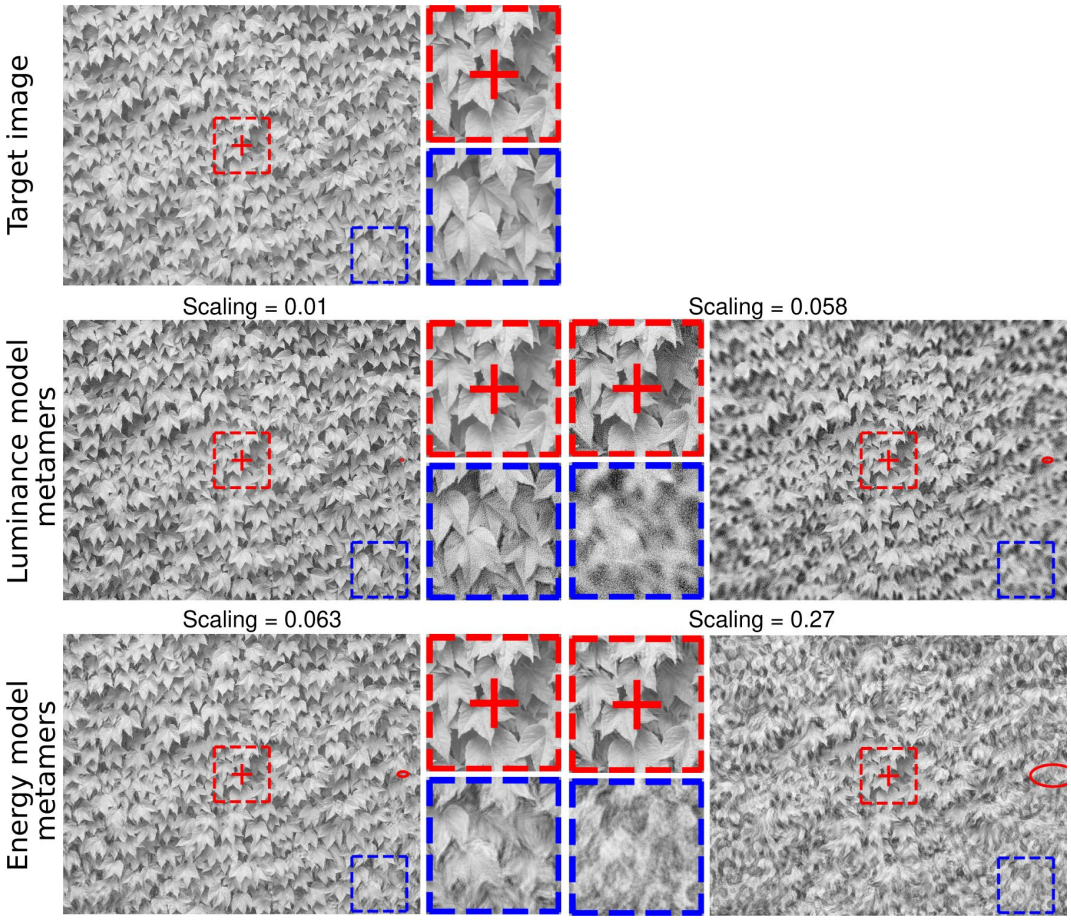


Figure 4.

Example synthesized model metamers. **Top:** Target image. **Middle:** Luminance model metamers, computed for two different scaling values (values as indicated, red ellipses to right of fixation indicate pooling window contours at half-max at that eccentricity). The left image is computed with a small scaling value, and is a perceptual metamer for most subjects: when fixating at the cross in the center of the image, the two images appear perceptually identical to the target image. Note, however, that when fixating in the periphery (e.g., the blue box), one can clearly see that the image differs from the target (see enlarged versions of the foveal and peripheral neighborhoods to right). The right image is computed with a larger scaling value, and is no longer a perceptual metamer (for any choice of observer fixation). **Bottom:** Energy model metamers. Again, the left image is computed with a small scaling value and is a perceptual metamer for most observers when fixated on the center cross. Peripheral content (e.g., blue box) contains more complex distortions, readily visible when viewed directly. The right image, computed with a large scaling value, differs in appearance from the target regardless of observer fixation. Full resolution version of this figure can be found on the OSF.

scaling value is approximately that associated with V1 receptive fields (***Freeman and Simoncelli, 2011)***. The high-scaling model metamer is perceptually distinct from the target image, and also perceptually distinct from the high-scaling luminance model metamer. The low-scaling model metamer, on the other hand, is difficult to distinguish from the original image (when fixating at the center), but is readily distinguished when one fixates peripherally.

The appearance of these two high-scaling metamers reflects the measurements that the models are matching and the seed images used for synthesis. The luminance model matches average pixel intensity, but has no constraints on spatial frequency, and thus its metamers retain the high frequencies present in the initial white-noise images. The energy model, on the other hand, matches the average contrast energy at all scales and orientations, but discards exact position information (which depends on phase structure). Hence, unlike the luminance model metamers, it reduces the high frequency power to match the typical content of natural images. Instead, it essentially scrambles the phase spectrum, leading to the cloud-like appearance of its metamers.

As can be seen in **figure 4**, both models can generate perceptual metamers. More generally, all pooling models can generate perceptual metamers if the scaling value is made sufficiently small (in the limit as scaling goes to zero, the model metamers must be identical to the target, in every pixel). For statistics that capture the relevant features for responses at some stage of the visual system, metamers can be achieved with windows whose size is matched to that of the underlying visual neurons. The maximal scaling at which synthetic images are perceptual metamers is thus highly dependent on the choice of underlying statistics: in our examples, the energy model perceptual metamer (**Fig. 4**, bottom left) is generated with a scaling value about six times larger than that for the luminance model perceptual metamer (middle left), and about five times smaller than those found in ***Freeman and Simoncelli (2011)*** using texture statistics. The goal of the present study is to use psychophysics to find the largest scaling value for which these two models generate perceptual metamers, known as the critical scaling.

Psychophysical experiment

We synthesized model metamers matching 20 different natural images (the target images) collected from the authors' personal collections, as well as from the UPenn Natural Image Database (***Tkačik et al., 2011)***, and an unpublished collection by David Brainard. The images were chosen to span a variety of natural image content types, including buildings, animals, and natural textures (**figure 5**). Model metamers were generated via gradient descent on the squared error between target and synthetic pooled statistics, and initialized with either an image of white noise or another image drawn from the set of target images.

In the experiments, observers discriminated two grayscale images, of size 53.6 by 42.2 degrees, sequentially displayed. Each image was presented for 200 msecs, separated by a 500 msec interval in which the screen was blank (mid gray). Each image was separated into two halves by a superimposed vertical bar (mid gray, 2 deg wide, see **figure 15**). One side, selected at random on each trial, was identical in the two images, while the other differed (e.g., one interval contains the target image, the other a synthesized model metamer). After the second image, a midgray screen appeared with text prompting the observer to report which side of the image had changed.

Critical scaling is four times smaller for the luminance than the energy model

We fit the behavioral data using the 2-parameter function introduced in ***Freeman and Simoncelli (2011)***, estimating the critical scaling (s_c) and maximum d' (a) parameters with a Markov Chain Monte Carlo procedure and a hierarchical, partial-pooling model similar to that used by ***Wallis et al. (2019)***.

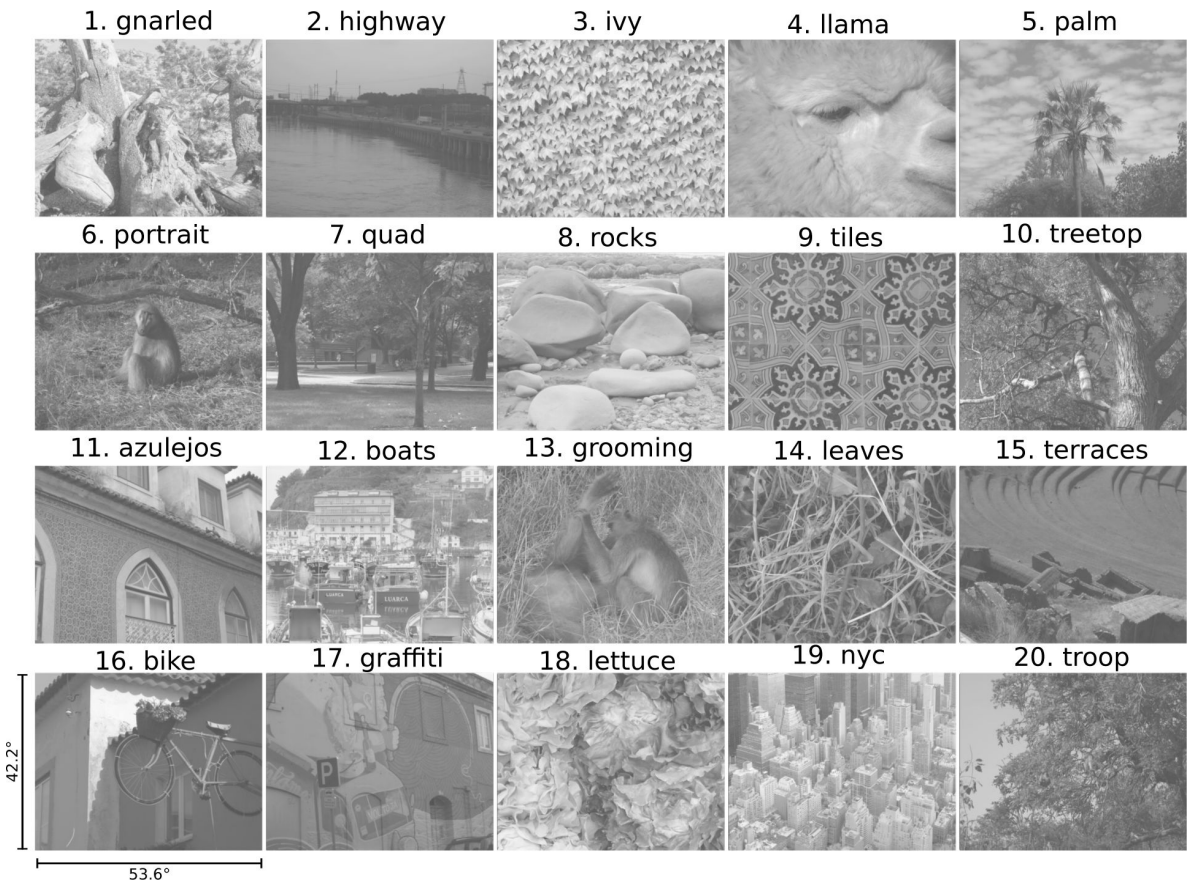


Figure 5.

Target images used in the experiments. Images contain a variety of content, including textures, objects, and scenes. All are relatively high-resolution RAW camera images, with values proportional to luminance and quantized to 16 bits. Images were converted to grayscale, cropped to 2048 x 2600 pixels, displayed at 53:6 x 42:2 degrees, with intensity values rescaled to lie within the range of [0.05, 0.95] of the display intensities. All subjects saw target images 1-10, half saw 11-15, and half saw 16-20. A full resolution version of this figure can be found on the OSF.

For a given model and comparison, performance increases monotonically with scaling, and is fit well by this particular psychometric function (**figure 6A**). The exception is the synthesized vs. synthesized comparison for the luminance model, where performance remains poor at all scales (see next section). In the original vs. synthesized cases (for both models), performance is near chance for the smallest tested scaling values tested and exceeds 90% for the largest. The critical scaling values, as seen in **figure 6B** are approximately 0.016 for the luminance model and 0.06 for the energy model. For comparison, we show the approximate scaling values for the receptive field diameters of retinal ganglion cells (both Midget, and Parasol), as well as for V1 cells (see appendix 2 for details on retinal ganglion cell scaling, and ***Freeman and Simoncelli (2011)*** for V1 scaling). The critical scaling for the luminance model falls between the two types of retinal ganglion cell. The critical scaling for the energy model is larger than that of both ganglion cell types, and approximately half of the lower end of the V1 range.

Critical scaling is lower for original vs. synthesized than synthesized vs. synthesized comparisons

For both luminance and energy models, it is generally easier to distinguish an original image from a synthesized image than to distinguish two synthesized images initialized with different white noise seeds (with same target image and scaling value). For the luminance model, discrimination of two synthesized images is nearly impossible at all scaling values. For the energy model, discriminating two synthesized images is possible but difficult, with performance only approaching 60%, on average (although note that there are substantial differences across subjects, see **figure 6B** and appendix 6). The critical scaling value for this comparison, 0.25, is close to the physiological scaling value estimated for V1, and comparable to that reported in ***Freeman and Simoncelli (2011)***. The asymptotic performance, however, is much lower in our data. We attribute this to experimental differences (see appendix section 4).

The difficulty of differentiating between two synthesized images is striking, as illustrated in **Figure 7**. In the limit of global pooling windows, luminance metamers are global samples of white noise which cannot be distinguished (***Wallis et al. (2019)*** made a similar point when discussing their use of the original vs. synthesized task). Analogously, synthesis with the energy model forces local orientated spectral energy to match, without explicitly constraining the phase. Two instances of phase scrambling within peripheral windows are not easily discriminable, even though either of the two might be discriminable from an image with more structure.

The interaction between model sensitivities and image content affects performance

To the extent that the models capture something important about human perception, image pairs that are model metamers will be perceptual metamers, and hence discrimination should be at chance. Neither model offers predictions of perceptual discriminability (they are deterministic, and do not specify any method of decoding or comparing stimuli). Consistent with this, the critical scaling, which measures the point at which image pairs become indistinguishable, does not vary much across images for a given model and comparison, unlike performance at super-threshold scaling values and the asymptotic levels of d^t (**figure 6B**). Variations in max d^t are especially clear in the images-specific psychometric functions for the original vs. synthesized energy model comparison (**figure 8**). Specifically, for the llama image, performance only rises slightly above chance, even at very large scaling windows. The respective target images in panel B suggest an explanation: much of the llama image is cloud-like pink noise, while the nyc image is full of sharp edges in the cardinal directions, with arise from precise alignment of phases across positions and scales. As discussed above, synthetic energy model metamers have matching local oriented spectral energy, with randomized phase information; in order to generate sharp, elongated contours for the buildings of nyc, the windows must be very small. Conversely, the appearance of

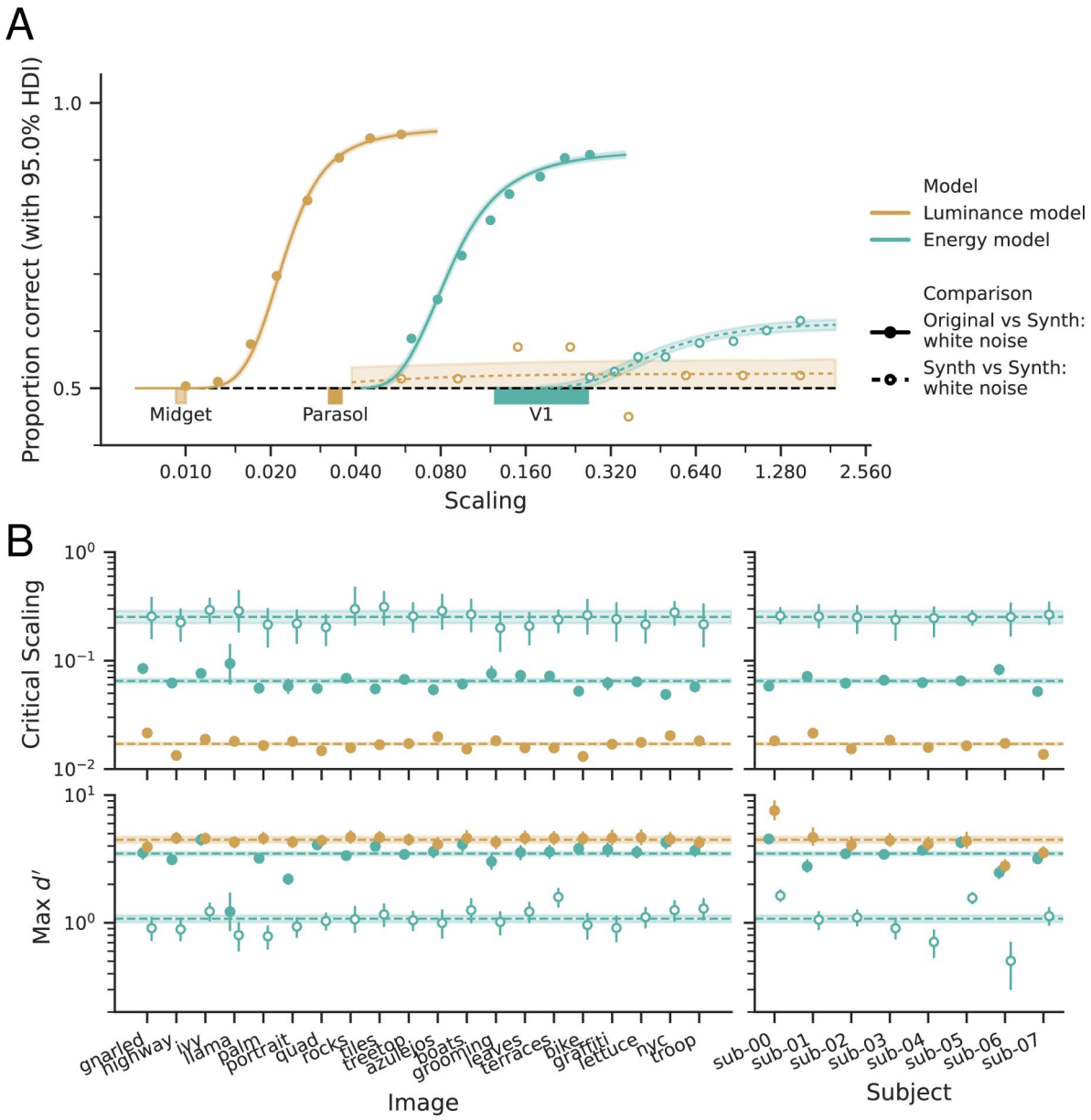


Figure 6.

Performance and psychophysical curve parameters values values for different models and image comparisons. The luminance model has a substantially smaller critical scaling than the energy model, and original vs. synthesized comparisons yield smaller critical scaling values than synthesized vs. synthesized comparisons. (A) Psychometric functions, expressing probability correct as a function of scaling parameter, for both energy and luminance models (aqua and beige, respectively), and original vs. synthesized (solid line) and synthesized vs. synthesized (dashed line) comparisons. Data points represent average values across subjects and images, 4320 trials per data point except for luminance model synthesized vs. synthesized comparison, which have only 180 trials per data point (one subject, five images). Lines represent the posterior predictive means of fitted curves across subjects and images, with the shaded region indicating the 95% high-density interval (HDI, ***Kruschke (2015)***). Horizontal bars (below dashed line at 0.5) indicate the range of physiological scaling values for the associated retinal ganglion cell type or cortical area. (B) Estimated parameter values, separated by image (left) or subject (right). Top row shows the critical scaling value and the bottom the value of the maximum *d'* parameter. Points represent the posterior means, shaded regions the 95% HDI, and horizontal dashed lines and shaded regions the global means and 95% HDI. Note that the luminance model, synthesized vs. synthesized comparison is not shown, because the data are poorly fit (panel A, beige dashed line).

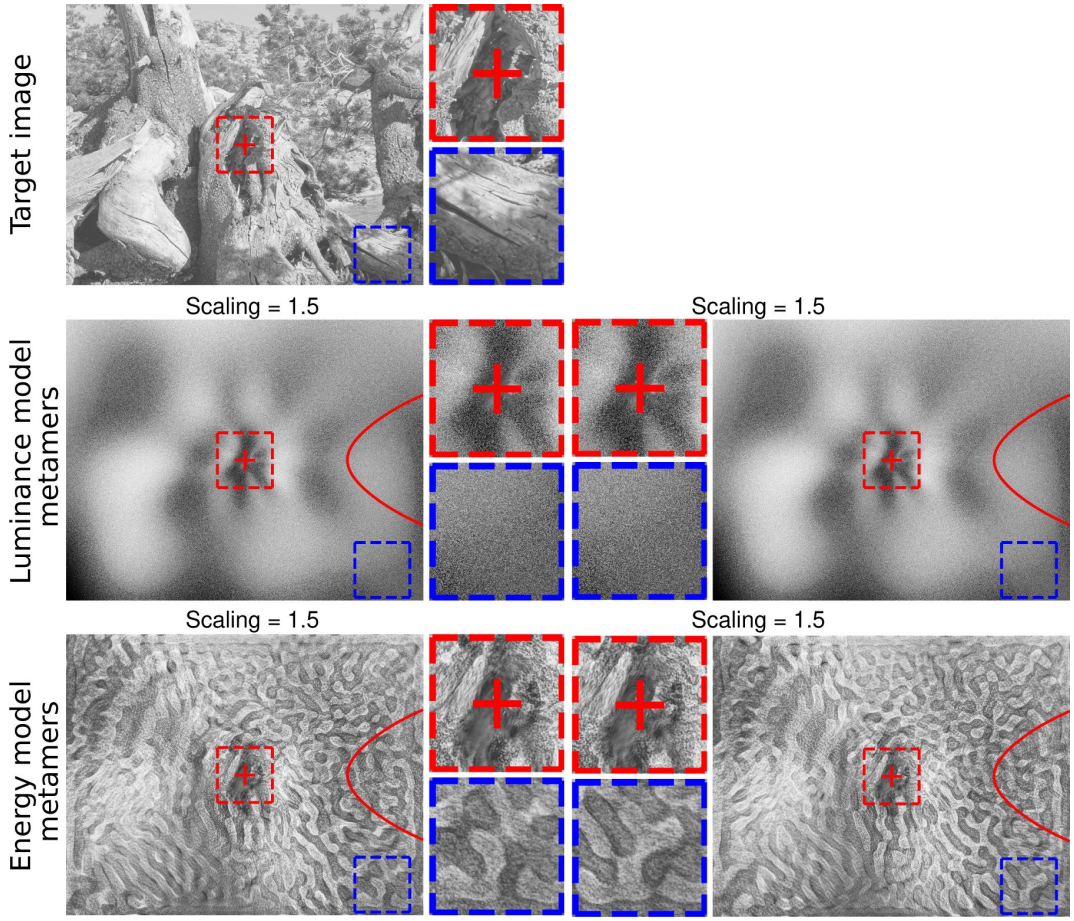


Figure 7.

Comparison of two synthesized metamers is more difficult than comparison of a synthesized metamer with the original image. For the highest tested scaling value (1.5) the original vs. synthesized comparison is trivial while the synthesized vs. synthesized comparison is difficult (energy model) or impossible (luminance model). **Top:** target image. **Middle:** Two luminance model metamers, generated from different initial uniform noise images. **Bottom:** Two energy model metamers, generated from different initial uniform noise images. All four of the model metamers can be easily distinguished from the natural image at top (original vs. synthesized), but are difficult to distinguish from each other, despite the fact that their pooling windows have grown very large (synthesized vs. synthesized). Full resolution version of this figure can be found on the OSF.

the llama is captured even when the pooling windows are large. Thus, when scaling is larger than critical scaling, some comparisons become easy and some do not. However, this pattern depends on the interaction between the model's sensitivities and the target image content, see appendix 6 for more details.

When discriminating two synthesized images, the initial image affects performance

The two types of comparisons shown in **figure 6** — original vs. synthesized and synthesized vs. synthesized — show very different critical scaling values. This indicates that for a particular scaling value and set of image statistics, some image pairs are much easier to discriminate than others. We hypothesize that metamers synthesized from white noise seeds are restricted to a relatively small region of the full set of model metamers. As a result, these images are more perceptually similar to each other than they are to the target image. To generate metamers outside of this set, we also used other natural images from our data set to initialize the synthesis procedure (which was not done in previous studies, ***Freeman and Simoncelli (2011)***; ***Wallis et al. (2019)***).

Figure 9A shows behavior for these additional comparisons in a single subject, sub-00. Changing the initialization image has a large effect on the synthesized vs. synthesized comparison but little-to-no effect on the original vs. synthesized comparison. For synthesized vs. synthesized, initializing with a different natural image improves performance compared to initializing with white noise, but is still worse than performance for original vs. synthesized. Importantly, these results cannot be predicted from the model, which gives no specific insight as to why some pairs are more discriminable than others.

Figure 9B shows three comparisons involving five metamers arising from different initializations, each with scaling corresponding to the vertical line in panel A, but with dramatically different human performance. The top row shows the easiest comparison, between the original image and a synthesized image initialized with a different natural image (bike); the subject was able to distinguish these two images with near-perfect accuracy (in this case, when comparing against a natural image, performance is identical regardless of whether the metamer was initialized with white noise or natural image). The bottom row shows the hardest comparison, between two synthesized images initialized with different samples of white noise. As discussed above, comparing two images of this type is difficult even with large pooling windows; at this scaling level, humans are insensitive to the differences between them, and so performance was at chance. The middle row shows two synthesized images, initialized with different natural images, which the subject was able to distinguish with moderate accuracy. When comparing these two images, one can see features in the periphery that remain from the initial image (tiles and highway, respectively). Even when fixating, the subject was able to use these features to distinguish the two images, i.e., the human was sensitive to them while the model was not. This reinforces the notion that the initialization of the synthesis process matters. In both the middle and bottom row, both images are synthesized (i.e., neither row contains the target image,) yet one comparison is much harder than the other.

The models reach critical scaling at different dimensionality

For each model, the number of statistics is proportional to the number of pooling regions, and thus decreases quadratically with scaling. **Table 1** shows average critical scaling values across all conditions, along with the corresponding number of model statistics. We can see that critical scaling does not correspond to a fixed number of statistics. We should also note that if one were to use the model outputs as a compressed representation of the image, the number of statistics in each representation is almost certainly an overcount, for several reasons. First, in order to ensure

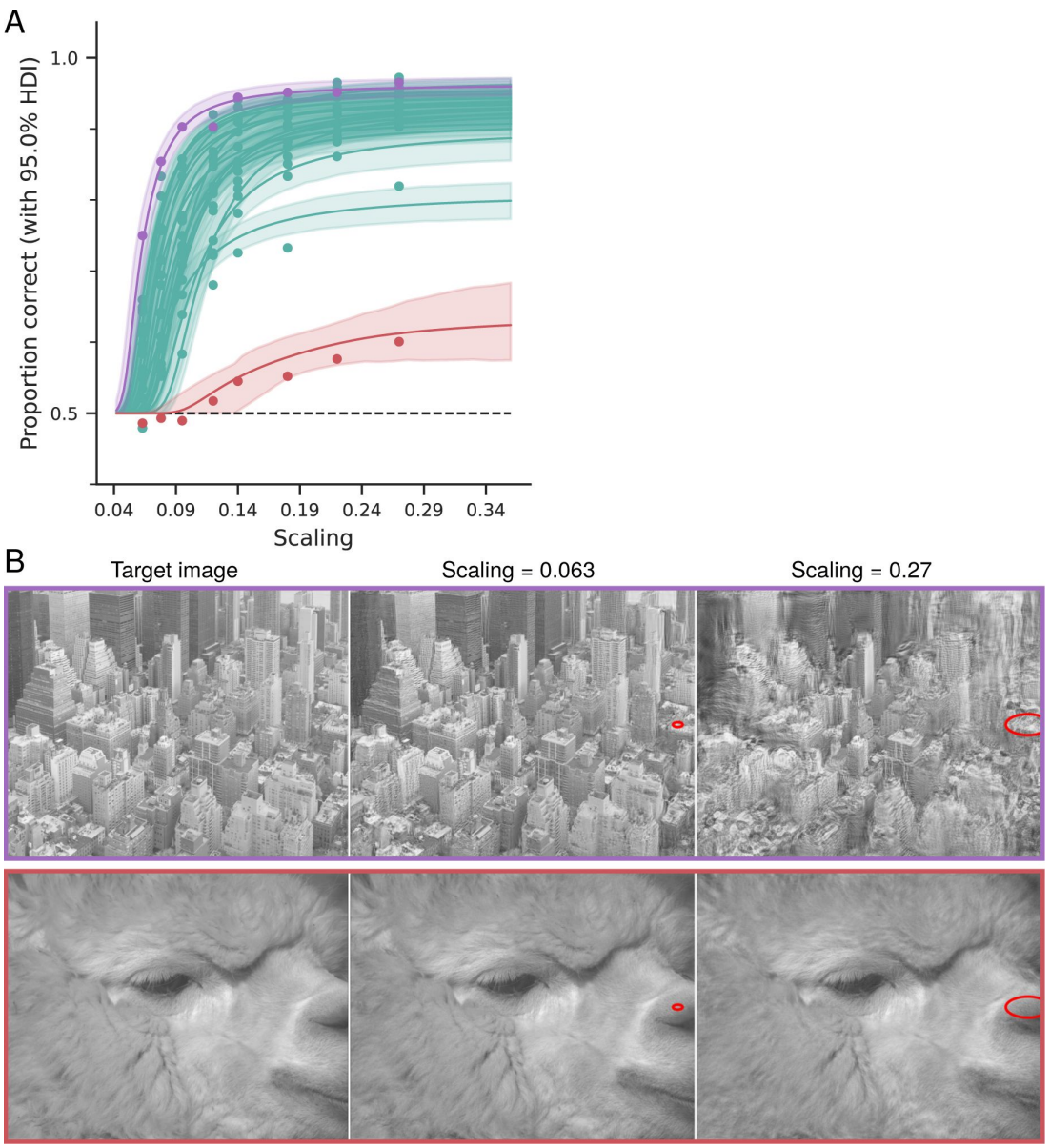


Figure 8.

The interaction between image content and model sensitivities greatly affects asymptotic performance, most noticeably on the synthesized vs. synthesized comparison for the energy model, while critical scaling does not vary as much. (A) Performance for each image, averaged across subjects, comparing synthesized images to natural images. Most images show similar performance, with one obvious outlier whose performance never rises above 60%. Data points represent the average across subjects, 288 trials per data point for half the images, 144 per data point for the other half. Lines represent the posterior predictive means across subjects, with the shaded region giving the 95% HDI. (B) Example model metamers for two extreme images. The top row (nyc) is the image with the best performance (purple line in panel A), while the bottom row (llama) has the worst performance (red line in panel A). In each row, the leftmost image is the target image, and the next two show model metamers with the lowest and highest tested scaling values for this comparison. Performance on the llama image is poor because much of the image content resembles pink noise. Thus, even with larger scaling values, the model metamers are very difficult to distinguish from the target image. The nyc image, on the other hand, contains hard edges with precise alignment of phase across scales. As the energy model discards phase information, this phase structure is lost in the model metamers, which are consequently easy to distinguish from the target image at all tested scaling values. However, this pattern does not hold in the luminance model, or for synthesized vs. synthesized comparisons, for which both images exhibit typical performance (see appendix figure 6). Full resolution version of this figure can be found on the OSF

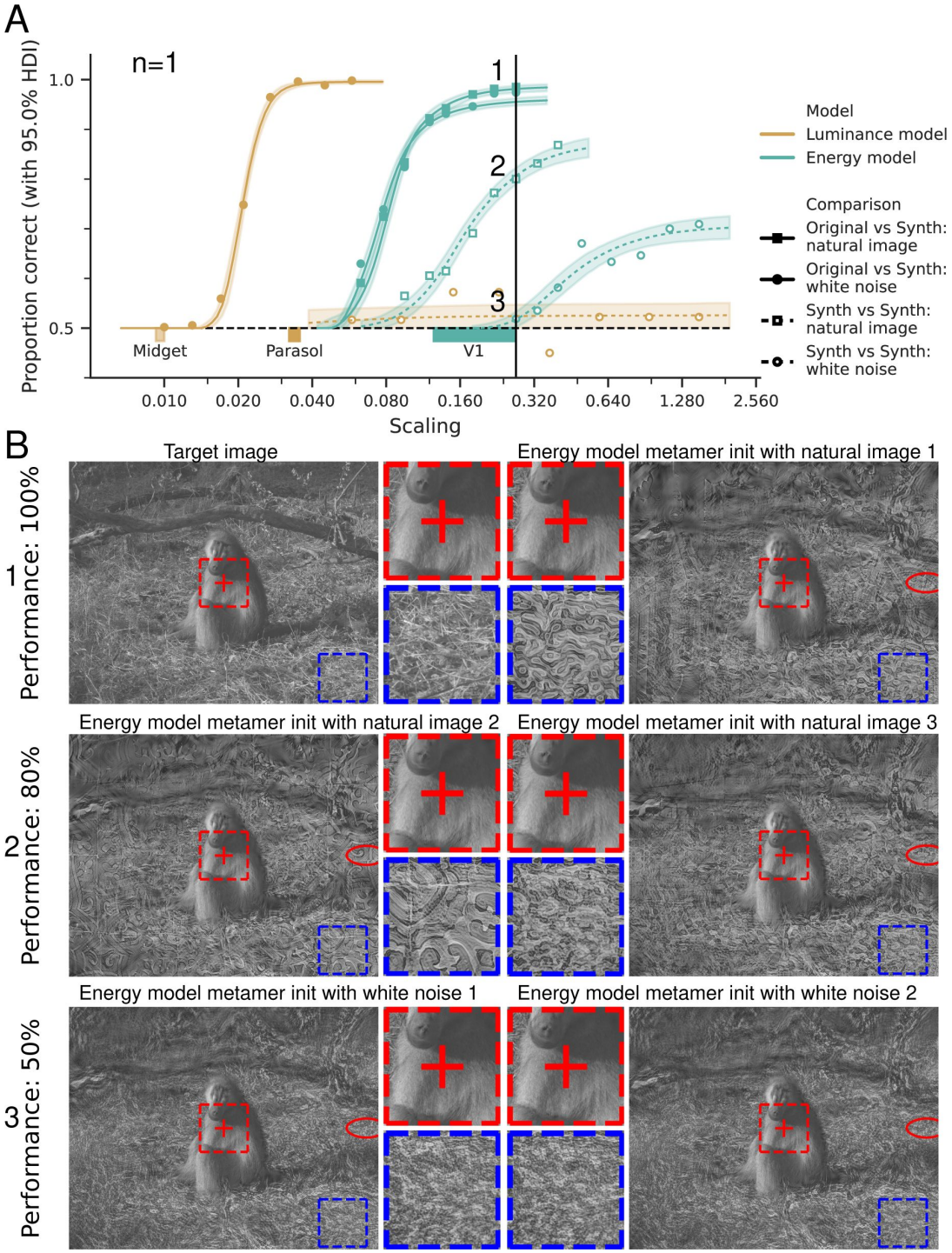


Figure 9.

Initializing model metamers with natural images does not affect performance in the original vs. synthesized comparison, but reduces critical scaling and increases max *d'* for the synthesized vs. synthesized comparison. Note all data in this figure is for a single subject and 15 of the 20 target images. (A) Probability correct for one subject (sub-00), as a function of scaling. Each point represents the average of 540 trials (over all fifteen images), except for the synthesized vs. synthesized luminance model white noise comparison (averaged over 5 images). Vertical black line indicates scaling value where difficulty ran from chance to 100%, based on initialization and comparison, as discussed in panel B. (B) Three comparisons corresponding to the three psychophysical curves intersected by the vertical black line in panel A. See text for details. Full resolution version of this figure can be found on the OSF.

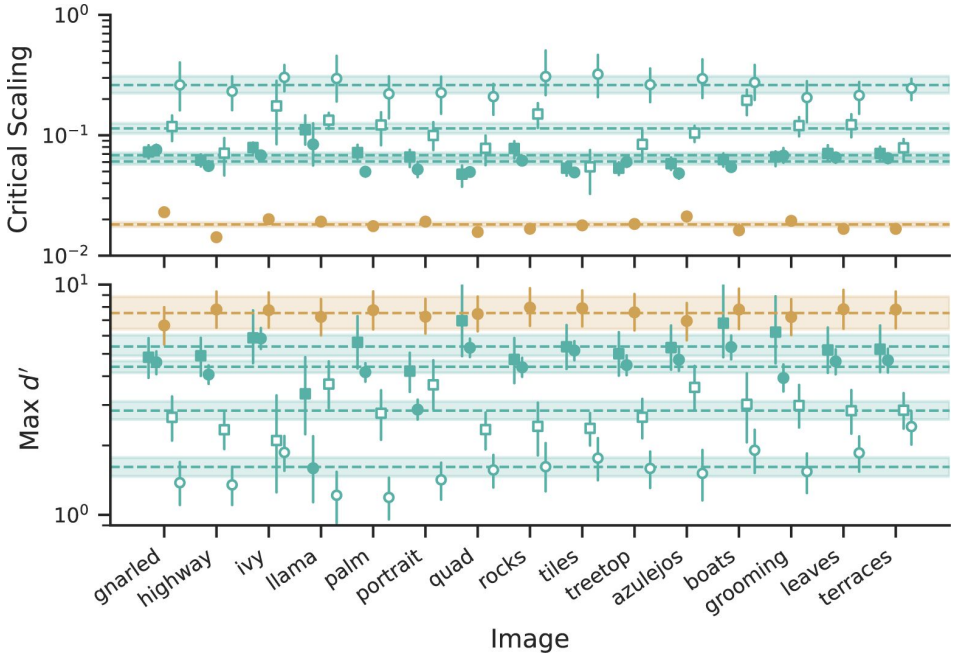


Figure 10.

Parameter values for the comparisons shown in **figure 9A** (Top: critical scaling value; Bottom: max *d'*). Data shown is from the single subject who completed all comparisons. Points represent the posterior means, shaded regions the 95% HDI, and horizontal dashed lines and shaded regions average across all shown images for this subject. Note that the luminance model, synthesized vs. synthesized: white noise comparison is not shown in this panel, because the data was poorly fit by this curve.

that the Gaussian pooling windows uniformly tile the image, the most peripheral windows in the model have the majority of their mass off the image, which is necessary to avoid synthesis artifacts. Second, for the energy model, we did not attempt to determine how the precise number of scales or orientations affected metamer synthesis, and currently all scales are equally-weighted across the image. As the human visual system is insensitive to high frequencies in the periphery and low frequencies in the fovea, this is probably unnecessary, and so some of these statistics can likely be discarded. Finally, our pooling windows are highly overlapping and thus the pooled statistics are far from independent; this redundancy means that the effective dimensionality of our model representations is lower than the quoted number of statistics.

Discussion

We measured perceptual discriminability of wide field-of-view metamers of two foveated pooling models of human vision. We found that performance depended on the model type (i.e., the type of image statistics that are pooled), the nature of the comparison (original vs. synthesized, or synthesized vs. synthesized), the seed image used for synthesis, and to a modest extent, the natural image target. Specifically, we found that critical scaling was much smaller for the luminance than for the energy model, and much smaller for original vs. synthesized than for synthesized vs. synthesized comparisons. For the former, we also found smaller critical scaling values when the synthesis was initialized from another natural image rather than white noise.

The linking proposition underlying the metamer paradigm

In ***Freeman and Simoncelli (2011)***, the authors propose a link between perceptual pooling models of the type found in this study and visual system physiology. They hypothesize that the critical scaling reflects the receptive field sizes of visual areas representing the corresponding statistics. Using this logic, they found a critical scaling for their texture model of approximately 0.5, and interpreted this as evidence of texture representation in V2, which has receptive field sizes with similar scaling. This logic follows the “Converse Identity” linking proposition in the framework proposed by

Teller (1984): identical perceptual states imply identical physiological states (at some stage of visual processing, and thereafter). Model-generated metamers provide an accessible experimental extension of this logic: at the critical scaling value, identical model outputs imply identical perceptual states, which imply identical physiological states. However, when attempting to link the critical scaling of a pooling model to the physiological scaling of a corresponding brain area, it is important to remember that receptive field size is a function of not just the visual area, but is also influenced by at least the cell class, cortical layer, mapping method, and stimulus type. Therefore, a visual area cannot be fully characterized by a single scaling value. This complicates the task of linking the psychophysical scaling value to the neural scaling value at a particular stage of visual processing. Hence we focus here on how the results vary with the type of model and type of comparison, but we do not seek to unambiguously assign a model to a stage of visual processing.

Visual computation: cascades of feature extraction and local pooling

Visual representations are formed through a cascade of transformations. An appealing hypothesis is that each of these is comprised of the same canonical “feature extraction and blur” computation, differing only in what is extracted and the spatial extent of the blurring (***Fukushima, 1980***; ***Douglas et al., 1989***; ***LeCun et al., 1989***; ***Heeger et al., 1996***; ***Riesenhuber and Poggio, 1999***; ***Bruna and Mallat, 2013)***. In the perception literature, ***Lettvin (1976)*** provides an early, informal discussion of this “compulsory feature averaging”, non-foveated versions have

Model	Comparison	Critical Scaling	Number of Statistics (percentage of image pixels)
Luminance	Original vs. Synth: white noise	0.017	19.6 %
	Synth vs. Synth: white noise	N/A	N/A
Energy	Original vs. Synth: white noise	0.065	34.7 %
	Original vs. Synth: natural image	0.068	31.5 %
	Synth vs. Synth: natural image	0.114	11.6 %
	Synth vs. Synth: white noise	0.252	2.6%

Table 1.

Critical scaling (posterior mean over all subjects and images) and number of statistics (as a percentage of number of input image pixels - a type of "compression rate"), for each model and comparison. Note that the critical scaling for original vs. synthesized comparisons does not result in the same number of statistics across models and, in particular, at their critical scaling values, all models have dimensionality less than that of the input image.

been described in ***Parkes et al. (2001)***; ***Pelli et al. (2004)***; ***Greenwood et al. (2009)***, and foveated proposals appear in ***Balas et al. (2009)*** and ***Freeman and Simoncelli (2011)***. In these, as in the current article, the stages are distinguished by their features, and the scaling of the pooling regions with eccentricity. The optics and photoreceptors pool the incident light over a small regions, V1 pools spectral energy over larger regions, V2 pools texture-like statistics over yet larger regions, and so forth. Consistent with this view, we find that for synthetic metamer stimuli, the pooled model statistic has a dramatic effect on the critical scaling value, which is approximately four times larger for the energy model than for the luminance model in the original vs. synthesized comparison (**figure 6** and **table 1**). This result is consistent across all subjects and all target images (**figure 6B**). The relationship between critical scaling and image statistic can be further appreciated by comparing these results to the model of ***Wallis et al. (2019)***, which was also fit to original vs. synthesized discrimination data: the critical scaling of our average energy model is about three times smaller than the average value for their texture model. Together, the three results show critical scaling ratios of approximately 1:4:12 for features corresponding to luminance:spectral-energy:texture, respectively (solid circles in **figure 11**).

Why does critical scaling depend on the comparison being performed?

We found large effects of comparison type on performance, consistent with those reported in ***Wallis et al. (2019)*** and ***Deza et al. (2019)***. For the energy model, with synthesis initialized from white noise, the critical scaling for synthesized vs. synthesized was about four times larger than that for original vs. synthesized. The maximum discriminability was also much lower for the former than the latter. For the luminance model, the difference was even more dramatic: participants were generally unable to discriminate any pairs for the synthesized vs. synthesized comparison. Differences between the two types of comparisons, which are summarized in **figure 11**, arise from an interaction between human perception, the model, and the synthesis process. Here we consider two abstract scenarios, one in which the comparison does matter and one in which it does not, in order to illustrate this interaction.

In the left panel of **figure 12**, the comparison does not matter: model metamers are distinguishable from each other *if and only if* they are distinguishable from the target image. The right panel illustrates a configuration in which this condition does not hold: two synthetic images can be indistinguishable, even though each are distinguishable from the target image. The idealized version implicitly assumes that the synthesized images sample the manifold of possible model metamers broadly, but our synthesis procedure (similar to those of ***Freeman and Simoncelli (2011)*** and ***Wallis et al. (2019)***) does not guarantee this. Initialization with natural images provides an intuitive, but ad-hoc, method of sampling from a broader portion of the manifold of possible model metamers, and results in smaller critical scaling values. More principled statistical sampling approaches (e.g., Markov Chain Monte Carlo) could result in more representative metamers.

Why does this critical scaling discrepancy decrease with increasing feature complexity?

As seen in **figure 11**, the difference in critical scaling between the two types of comparisons declines as the image statistics being pooled become more complex. While the original vs. synthesized critical scaling value is lower for all three models, the gap between the two decreases as the models increase in complexity: infinite for the luminance model, roughly quadruple for energy, and less than double for texture (see **figure 11**).

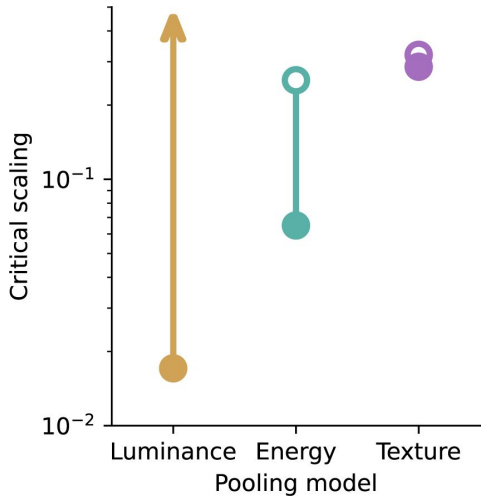


Figure 11.

Critical scaling values for the two pooling models presented in this paper (Luminance and Energy) and the Texture model (originally tested in ***Freeman and Simoncelli (2011)***, data from ***Wallis et al. (2019)***, averaging across the two image classes). Solid points indicate the original vs. synthesized white noise comparisons, while hollow points indicate synthesized vs. synthesized white noise comparisons (for the luminance model, this is effectively infinite, since participants were unable to complete the task). For all three models, critical values are smaller in the original vs. synthesized comparison than the synthesized vs. synthesized one, and their ratio decreases with increasing complexity of image statistics. A potential explanation for this is that the more complex models approximate computations performed in deeper levels of the visual hierarchy, beyond which there are fewer remaining stages to discard information.

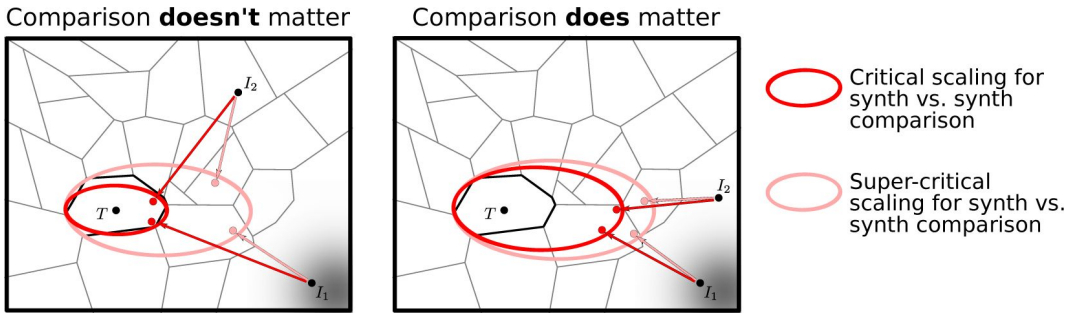


Figure 12.

In the idealized version of the metamer paradigm, synthesized images broadly sample the space of possible model metamers. Thus, at large scaling values, synthesized images are distinguishable from each other *and* the target image, and they become indistinguishable from each other at the same scaling value for which they are indistinguishable from the target (left panel). In this case, model metamers are metameric with each other if and only if they are metameric with the target image. In our experements, however, this is not the case: at some scaling values, two synthesized images can be metameric with each other, but distinguishable from the target image (right panel).

One potential explanation for this observation is that these computations are being performed in deeper stages of the visual hierarchy and there are progressively fewer opportunities to discard information later in the hierarchy. For example, the difference in V1 responses for a pair of images may be discarded by a later stage (e.g., area IT), but there are not many steps of processing between IT and the perceptual read out where differences between IT responses can be discarded. This may explain why we see no overlap between the critical scaling values for original vs. synthesized and synthesized vs. synthesized comparisons across images in **figure 6B**, whereas ***Wallis et al. (2019)*** find substantial overlap for the texture model. Ultimately, these possibilities can only be distinguished through further physiological measurements.

Another potential explanation is schematized in **figure 13**: for the luminance model, the model metamer classes are aligned with the perceptual metamer classes and the white noise samples such that, for a given scaling value, all model metamers initialized with white noise fall into the same perceptual metamer class. This is an extreme case of the synthesis issue described in the previous section. On the other hand, the texture model's metamer classes are orthogonal to the white noise samples, so that synthesized model metamers easily fall into distinct perceptual metamer classes, leading to a critical scaling value that is much more similar to the value for the original vs. synth comparison.

Why and when does the critical scaling depend on synthesis initialization?

The previous sections show that the critical scaling depends on the comparison type and feature complexity. We also find that it depends on synthesis initialization. Natural images are more likely than synthetic images to include information that the human visual system has evolved to discriminate, as opposed to information that is discarded at some later stage of processing. Using natural images to initialize synthesis (or development of novel synthesis methods to better explore the space of metamers) may reduce the discrepancy between the two conditions.

The schematic presented in **figure 14** provides a potential explanation for why critical scaling values depend on the comparison and synthesis initialization. Ideally, synthesized metamers are discriminable from each other if and only if they are also discriminable from the target, as described in section Why does critical scaling depend on the comparison being performed?. The variability in critical scaling across comparisons indicates that this condition is being violated: the synthesized vs. synthesized conditions have a higher critical scaling value indicates that the synthesized images used in those comparisons are falling into the same metamer class. This effect is reduced when initializing with a natural image.

This suggested the noise-initialized algorithm produces a biased sampling of the space of all possible model metamers. For each target image, model, scaling value, and initialization condition (white noise or natural image), we generated three model metamers. If these were distributed across the space of all possible model metamers, we would not see the dependence on initialization depicted in **figure 14**. Initializing with a natural image was one attempt to sample a different portion of this space, and it did reduce the discrepancy between comparison conditions. However, further work needs to be done to better understand the effects of initialization on generated samples. One possibility that seems promising would be to synthesize model metamers for a given target image, model, and scaling value in sets that are as different from each other as possible, quantified using different pooling models, other visual models, or image quality metrics. We believe the metamer paradigm can be made more informative by paying more attention to the synthesis procedure.

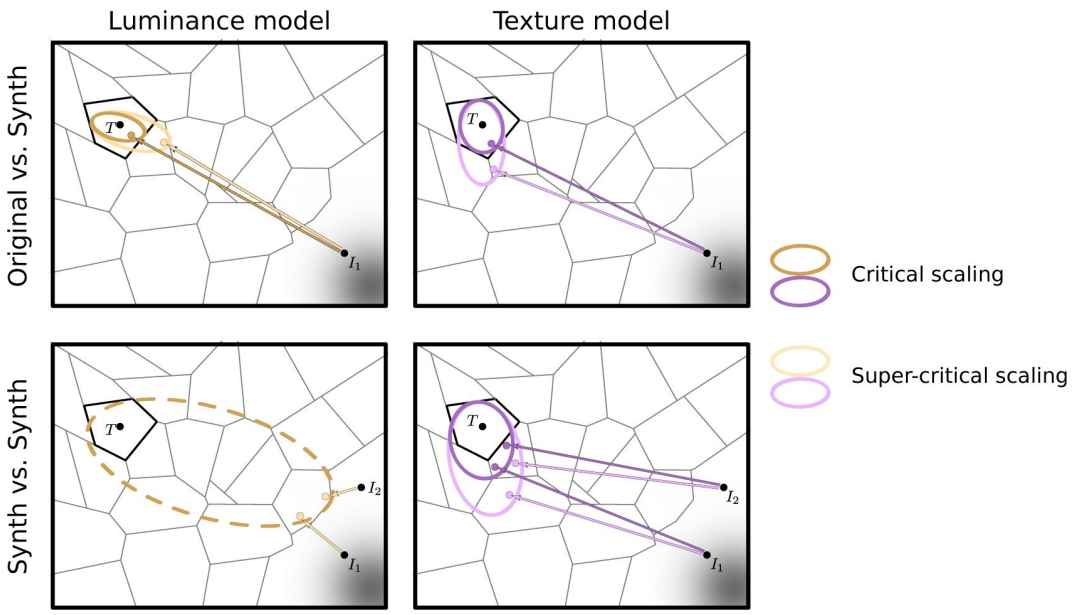


Figure 13.

The difference between the synth vs. synth and original vs. synth white noise comparisons decreases as model complexity increases (compare with **figure 12**). Because of the alignment between the luminance model metamer classes, the underlying perceptual metamer classes, and the white noise samples used to initialize the synthesis process, synthesized images *always* lie within the same perceptual metamer class and are thus never distinguishable. On the other hand, the texture model's metamer classes grow orthogonally to the white noise samples, resulting in a much smaller critical scaling value for the synth vs. synth white noise comparison.

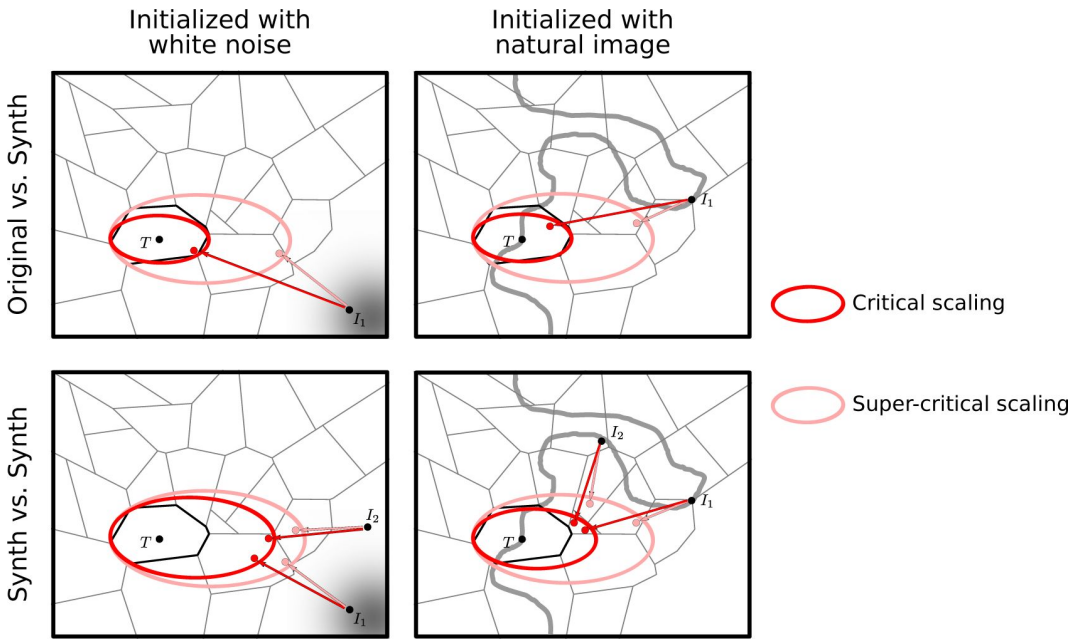


Figure 14.

Schematics describing results presented in this paper. Unlike the idealized metamer paradigm (see section Why does critical scaling depend on the comparison being performed? and **figure 12A**), the critical scaling value for our models depends on the comparison being performed and, to a lesser extent, the image used to initialize synthesis. For the original vs. synthesized comparison, this does not affect the critical scaling value, as the synthesized images are always distinct from the target. However, initializing with white noise can result in synthesized images that lie in the same metamer class as each other even while they are distinct from the target, resulting in a relatively large scaling value for the synthesized vs. synthesized comparison. Initializing with natural images reduces the magnitude of this phenomenon.

We believe a more extreme version of this situation is reflected in the data for the luminance model: the synth vs. synth comparison is *never* possible when initializing model metamers with white noise (so we cannot estimate the model's critical scaling), but initializing with natural images does allow these model metamers to be distinguished from each other (see appendix 1). We believe this reflects an extreme version of the biased sampling discussed above: the local luminance matching enforced by the luminance model is a fairly lax constraint and thus initializing with white noise leads to model metamers that consistently lie within the same perceptual metamer class, while at the same time being very discriminable from the target image (see **figure 7**).

By gathering discriminability data for different types of comparisons, we are able to get a better sense of how the metamer classes of our pooling models align with those of the visual system. Our current data supports the proposal from ***Wallis et al. (2019)*** that the critical scaling should be estimated using only the original vs. synthesized comparison, as those values are minimal across comparisons and do not seem to depend on how metamer synthesis is initialized. Furthermore, there are many applications where that is the only comparison that matters (e.g., for generating images that must be indistinguishable from natural images only). However, it is not clear that this will always be the case: while we never sampled from the metamer class containing the target image while the critical scaling value was large, it is theoretically possible, and additional comparisons can provide more protection against over-interpretation results stemming from biased sampling.

Asymptotic performance, but not critical scaling, depends on image content

Similar to ***Wallis et al. (2019)***; ***Brown et al. (2021)***, we find that metamer discrimination performance is somewhat dependent on image content. Both of those studies synthesize model metamers based on pooled texture statistics, and ***Wallis et al. (2019)*** shows that texture-like original images are harder to distinguish from their synthesized images than scene-like ones, while ***Brown et al. (2021)*** show that original textures with higher global and local regularity (e.g., woven baskets) are the easiest to distinguish from their synthesized images than those with low regularity (e.g., animal fur). This aligns with our result: the most distinguishable pairs include natural images with features not well-captured by the synthesizing model, whereas the least distinguishable include those natural images whose features are all adequately captured.

However, we should note that we found this image-level variability largely in super-threshold performance, and this variability does not constitute a failure of these pooling models. As pointed out by ***Freeman and Simoncelli (2011)***, asymptotic performance also varies with experimental manipulation, while critical scaling remains relatively unaffected. The metamer paradigm makes strong predictions about what happens when the representation of two stimuli are matched: they are indistinguishable, and so performance on a discrimination task will be at chance, as captured by the critical scaling value. However, it makes *no* predictions about performance at super-threshold levels, as captured by the max d^t parameter. An analogy with color vision seems apt: color matching experiments provide evidence for what spectral distributions of light are perceived as identical colors, but provide no information about whether humans consider blue more similar to green or to red; further investigations are necessary to understand color appearance. Thus, while this image-level variability is worth investigating in order to better understand the sensitivities of our model, it does not much affect the inferences we want to make about the human visual system, and speaks more to the need for a complementary approach (see next section).

Observer models are needed to predict discriminability of non-metameric images

As discussed above, the metamer paradigm's converse identity linking proposition is silent on what is implied by *distinct* model outputs and so a complementary approach is required, such as building observer models to predict perceptual distances. Specifically, the metamer paradigm makes no predictions about discriminability for model metamers synthesized with a super-critical scaling value. Such images are discriminable to the synthesizing pooling model at critical scaling. However, the information distinguishing them might be discarded by later stages of visual processing (center panel, **figure 1**). A complementary approach investigating how such differences are handled by later brain areas is necessary to gain a better understanding of image discriminability beyond "identical or not". Such work could use the models presented here as a starting point and could draw on the substantial literature of observer models in vision science and image processing.

There are several important properties of the pooling models used in metamer studies that should be revisited if attempting to extend them into observer models. First, the models assume equal sensitivity to all statistics, and that the sensitivity to these statistics does not change across the visual field. Second, the models assume that every statistic is pulled in regions of equal size (e.g., that high frequency is pooled over the same size region as low frequency). Finally, the shape of the pooling windows (e.g., whether windows should be radially elongated with a 2:1 ratio) and their scaling with eccentricity can probably benefit from refinement. When performing the task (especially the original vs. synthesized comparisons), subjects reported that the most informative portions of the image were in the mid-periphery, rather than close to fixation or at the edges of the image. This suggests that the windows may be too large in the mid-periphery, and that window width may be better modeled in a non-linear manner.

The investigation of super-threshold performance and appearance is a necessary complement to the metamer paradigm, which focuses on the question of whether two images are perceptually identical or not. Whereas observer models and image quality metrics often rely on natural images, common experimental stimuli, or sets of common distortions, the metamer paradigm relies on the synthesis of novel images, often turning up unexpected exemplars. Combining and extending this synthesis-focused approach with an attention to super-threshold performance would help lead to a fuller understanding of human perceptual sensitivities and insensitivities.

Materials and Methods

All experimental materials, data, and code for this project are available online under the MIT or similarly permissive licenses. Specifically, software is on GitHub, synthesized metamers can be browsed on this website, and all images and data can be downloaded from the OSF. The GitHub site provides instructions for downloading and using data.

Synthesis

We synthesized model metamers matching 20 different natural images (the target images) from the authors' (W.F.B and E.P.S) personal collections, as well as from the UPenn Natural Image Database (***Tkačik et al., 2011)*** and from an unpublished collection by David Brainard. The selected photos were high-resolution with 16-bit pixel intensities proportional to luminance, that had not undergone lossy compression (which could result in artifacts). They were converted to grayscale using scikit-image's color.rgb2gray function (***van der Walt et al., 2014)***, cropped to 2048 by 2600 pixels (the Brainard photos were 2014 pixels tall, so a small amount of reflection padding was used to reach 2048 pixels), and had their pixel values rescaled to lie between 0.05 and 0.95. Synthesized images were still allowed to have pixel values between 0 and 1; without

rescaling the target images, synthesis resulted in strange artifacts with pixels near 0, as this was the minimum allowed value. The images were chosen to span a variety of natural image content types, including buildings, animals, and natural textures (see **figure 5**).

We synthesized the model metamers using custom software written in PyTorch (***Paszke et al., 2019)***, using the AMSGrad variant of the Adam optimization algorithm (***Kingma and Ba, 2014***; ***Reddi et al., 2018)***, with learning rate 0.01. Slightly different approaches were used for the luminance and energy model metamers. For the luminance model metamers, the objective function was to minimize the mean-squared error between the model representation of the target and synthesized images, $L(x, \hat{x}) = (M(x) - M(\hat{x}))^2$, and synthesis was run for 5000 iterations. For the energy model metamers, the objective function also contained a quadratic range penalty term, which penalized any pixel values outside of [0, 1], $L(x, \hat{x}) = .5(M(x) - M(\hat{x}))^2 + .5B(\hat{x})$, where

$$B(\hat{x}) = \begin{cases} \hat{x}^2 & \hat{x} < 0 \\ 0 & \hat{x} \in [0, 1] \\ (\hat{x} - 1)^2 & \hat{x} > 1 \end{cases}$$

Synthesis was run for 15000 iterations. Additionally, energy model metamer synthesis used stochastic weight averaging (***Izmailov et al., 2018)***, which helped avoid local optima by averaging over pixel values as synthesis neared convergence, and used coarse-to-fine optimization (***Portilla and Simoncelli, 2000)***. Additionally, each statistic (in both models) was z-scored using the average statistic value computed across the entire image on a selection of grayscale texture images. For both models, synthesis terminated early if the loss had not decreased by more than $1e-9$ over the past 50 iterations. While not all model metamers achieved the same loss values, with differences in synthesis loss across target images, there was no relationship between the remaining loss and behavioral performance.

For each model, its windows were represented as two tensors, one for angular slices and one for annuli, which, when multiplied together, would give the individual windows, with separate sets of windows for each scale in the energy model. This required a large amount of memory, and so for scaling values below 0.09, models were too large to perform synthesis on the available NVIDIA GPUs with 32GB of memory. Thus, all luminance model metamers were computed on the CPU, and synthesis of a single image took from about an hour for scaling 1.5 to 2 days for scaling 0.058 to 14 days for scaling 0.01. For the energy model metamers, the lowest two scaling values were computed on the CPU, with synthesis taking about a week. For those energy model metamers which were able to be computed on the GPU, synthesis took from 5 hours for scaling 0.095 to 1.5 hours for scaling 0.27 and above. This synthesis procedure was completed in parallel using the high-performance computing cluster at the Flatiron Institute.

Synthesized images for original vs. synthesized and synthesized vs. synthesized white noise comparisons (see Psychophysical experiment) were initialized with full-field patches of white noise (each pixel sampled from a uniform distribution between 0 and 1). For each model, scaling value, and target image, three different initialization seeds were used. A unique set of three seeds was used for each scaling value and target image, except for the following, which all used seeds {0, 1, 2}:

- Luminance model: azulejos, bike, graffti, llama, terraces, tiles; scaling 0.01, 0.013, 0.017, 0.021, 0.027, 0.035, 0.045, 0.058, 0.075 and 0.5.
- Energy model: azulejos, bike, graffti, llama, terraces, tiles; scaling 0.095, 0.12, 0.14, 0.18, 0.22,0.27, 0.33, 0.4, and 0.5.

For original vs. synthesized and synthesized vs. synthesized natural image comparison, synthesized images for each model, scaling value, and target image were initialized with three random choices from among the rest of the target images.

Pooling windows

Pooling model windows are laid out in a log-polar grid, with peaks spaced one standard deviation apart, such that adjacent window functions cross at a value of 0.352 (relative to max of 0.4). They have a single parameter, scaling, which specifies the ratio of the pooling window diameter at full-width half-max in the radial direction and the distance of its center from the fovea, both in degrees. For example, the pooling windows of a model with scaling factor 0.1 have a radial diameter of 1 degree at 10 degrees eccentricity, 2 at 20 degrees, etc.

Image pixels within 0.5 degree from the fixation point were exactly matched in our synthesized images, approximating the fovea, where no pooling occurs. Additionally, for small scaling values, windows for some distance beyond this region would be smaller than a pixel and so the only solution is to match the pixel values in that region directly. For example, with image resolution of 2048 by 2600 and display size of 53.6 by 42.2 degrees, models with scaling value of 0.063 have windows whose diameter at FWHM is smaller than a pixel out to 0.52 degrees, with this number increasing quadratically as scaling decreases, reaching 3.29 degrees for scaling 0.01 (see 5 for more discussion).

Observers

Eight participants (5 women and 3 men, aged 24 to 33), including an author (W.F.B.), participated in the study and were recruited from New York University. All subjects had normal or corrected-to-normal vision. Each subject completed nine one-hour sessions. One subject (sub-00) also performed seven additional sessions. All subjects provided informed consent before participating in the study. The experiment was conducted in accordance with the Declaration of Helsinki and was approved by the New York University ethics committee on activities involving human subjects.

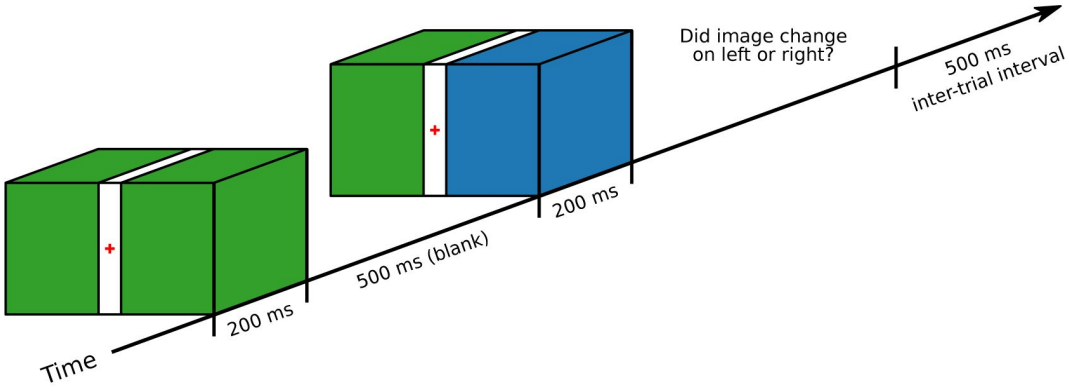
Psychophysical experiment

A psychophysical experiment was run in order to determine which of the synthesized model metamers were also perceptual metamers. We first describe the structure of a single trial, then how the trials were organized into blocks and sessions.

Trial structure

See **figure 15** for schematic. Observers viewed a series of grayscale images on a monitor, at a size of 53.6 by 42.2 degrees. An initial image, divided in half by a vertical midgray bar 2 degrees wide, was displayed for 200 msecs, before being replaced by a midgray screen for 500 msecs, followed by a second image for another 200 msecs (also divided by a vertical midgray bar). Images were presented for 200 msecs to minimize the possibility of eye movements. The dividing bar prevented participants' use of discontinuities between the two image halves to perform the task. One side of the second image (left half or right half) was identical to the first image, and the other side changed. After the second image was viewed, a midgray screen appeared with text prompting a response, and the observer's task was to report which half had changed. The observer had as much time as necessary to respond. The two compared images were either two synthesized images (synthesized for identical models with the same scaling value and target image, but different initializations) or one synthesized image and its target image. Either image could be presented first.

The midgray blank screen presented between the two image presentations reduces motion cues participants could use to discriminate the two images. Our models aim to capture the steady state response to the images, not the transient response. The mask forces the participants to use the image content to discriminate between the two images, rather than relying on temporal edges (analogous to our use of the vertical bar to prevent the use of spatial edges). This introduces a short-term memory component in the task (participants must remember the first image in order



	First image	Second image left	Second image right
Original vs. Synth	Original	Original	Synth
		Synth	Original
	Synth	Original	Synth
		Synth	Original
Synth vs. Synth	Synth *i*	Synth *i*	Synth *j*
		Synth *j*	Synth *i*
	Synth *j*	Synth *i*	Synth *j*
		Synth *j*	Synth *i*

Figure 15.

Schematic of psychophysics task. Top shows the structure for a single trial: a single image is presented for 200 msec, bisected in the center by a gray bar, followed by a blank screen for 500 msec. The image is re-displayed, with a random half of the image changed to the comparison image, for 200 msec. The participants then have as long as needed to say which half of the image changed, followed by a 500 msec intertrial interval. Bottom table shows possible comparisons. In original vs. synthesized, one image was the target image whose model representation the synthesized images match (see **figure 5**), and the other was one such synthesized image. In synthesized vs. synthesized, both were synthesized images targeting the same original image, with the same model scaling, but different initialization. In experiments, dividing bar, blanks, and backgrounds were all midgray. For more details see text.

to compare it to the second image), as in previous metamer discrimination experiments (***Freeman and Simoncelli, 2011***; ***Deza et al., 2019***; ***Wallis et al., 2019)***. We believe the precise duration of this mask is unimportant for our results: first, ***Bennett and Cortese (1996)*** found the duration of a blank screen did not affect thresholds in a spatial frequency discrimination task over a range from 200 to 10,000 msec, and second, mask duration is likely to have a similar effect on performance as image presentation duration, which ***Freeman and Simoncelli (2011)*** found affected asymptotic performance but not critical scaling.

Session and block organization

Across 9 sessions, each subject completed a total of 12,960 trials, factored into 3 model/comparison combinations by 8 scaling values by 15 target images by 36 repetitions. The 3 model/comparison combinations were 1) luminance model, original vs. synth, white noise; 2) energy model, original vs. synth, white noise; and 3) energy model, synth vs. synth, white noise. The 8 scaling values were logarithmically spaced, with the range chosen separately for each model/comparison to span an appropriate range of values. For luminance model, original vs. synth, white noise, the scaling endpoints were 0.01 and 0.058; for energy model, original vs. synth, white noise, the endpoints were 0.063 and 0.27; and energy model, synth vs. synth, white noise, the endpoints were 0.27 and 1.5. There were a total of 20 target images, but each subject only saw 15. Every subject saw images 1 through 10. Half the subjects also saw images 11 through 15, and half saw images 16 through 20 (see **figure 5**). The 36 repetitions were averaged for analysis and included 12 trials for each of 3 synthesis seeds. For the white noise-initialized comparisons, these seeds were independent samples of white noise used to initialize the synthesis procedure, resulting in 3 distinct model metamers.

Each of the above model/comparisons was tested across 3 sessions, each lasting approximately one hour. Each subject started with either the luminance or energy model, original vs. synth, white noise. The 3 sessions required for the model/comparison tested first were completed before moving onto 3 sessions testing the other model. The order of the two models was randomized across subjects. After completing these 6 sessions, the subject completed 3 sessions testing the energy model, synth vs. synth, white noise. This comparison was last as it was the most difficult.

Each of the 9 sessions consisted of 1,440 trials, containing all 36 repetitions for all 8 scaling values for 5 of the 15 target images viewed by the subject (target images were randomly assigned to sessions, independently for each subject). The 1,440 trials per session were broken up into 5 blocks of 288 trials each. Each block took about 8 to 12 minutes, and consisted of 12 repetitions for all 8 scaling values for 3 of the 5 target images.

In addition, one subject (sub-00) completed 7 additional sessions (10,080 additional trials). This included 1 session for luminance model, synth vs. synth, white noise; 3 for energy model, original vs. synth, natural image; and 3 for energy model, synth vs. synth, natural image. As with the comparisons that all subjects completed, these sessions each included 1,440 trials, factored into 5 target images by 8 scaling values by 36 repetitions. Only 1 session was included for luminance model, original vs. synth, white noise because performance was at chance for all images and all scaling values (see **figures 6** and **7**). No sessions were completed for the luminance model, natural image comparisons due to the time required for synthesis; see appendix section 1 for more information.

The four types of comparisons are explained in full below:

1. Original vs. synthesized, white noise: the two images being compared were always one synthesized image and its target image, and the synthesized image was initialized with a sample of white noise.

2. Synthesized vs. synthesized, white noise: both images were synthesized, with the same model, scaling value, and target image, but different white noise seeds as synthesis initialization.

3. Original vs. synthesized, natural image: the two images being compared were always one synthesized image and its target image, and the synthesized image was initialized with a different natural image drawn randomly from our set.

4. Synthesized vs. synthesized, natural image: both images were synthesized, with the same model, scaling value, and target image, but initialized with different natural images from our set.

Subjects completed several training blocks. Before their first session, they completed an initial training block, comparing two natural images and two noise samples (one white, one pink). Before their first session of each comparison type including a natural image, they completed a secondary training block showing two natural images and two synthesized images of the type included in the session, one with the largest scaling included in the task and one with the smallest. Before the session comparing two synthesized images, they similarly completed a training block comparing four synthesized images, two with a low scaling value and two with a high scaling value, for each of two target images. Each training block took one to two minutes and was repeated if performance on the high scaling synthesized images was below 90% or subjects expressed uncertainty about their ability to perform the task (participants were expected to perform close to chance for the low scaling synthesized images). Additionally, before each session which included a natural image (the original vs. synthesized comparisons), subjects were shown the five natural images that would be part of that session, as well as two example synthesized images per target image, one with a low scaling value, one with a high scaling value. Before each session comparing two synthesized images (the synthesized vs. synthesized comparison), subjects were shown four example synthesized images per target image, two with the lowest scaling value and two with the highest scaling value for that comparison. A video of a single energy model training block, original vs. synthesized: white noise comparison, can be found on the OSF.

Apparatus

The stimuli were displayed on an Eizo CS2740 LED flat monitor running at 60 Hz with resolution 3840×2160. The monitor was gamma-corrected to yield a linear relationship between luminance and pixel value. The maximum, minimum, and mean luminances were 147.73, .3939, and 77.31 cd/m^2, respectively.

The experiment was run with a viewing distance of 40 cm, giving 48.5 pixels per degree of visual angle. A chin and forehead rest was used to maintain head position, but the subjects' eyes were not tracked.

The experiment was run using custom code written in Python 3.7.0 using PsychoPy 3.1.5 (***Peirce et al., 2019)***, run on an Ubuntu 20.04 LTS desktop. A button box was used to record the psychophysical response data. All stimuli were presented as 8-bit grayscale images.

Data analysis

All trials were analyzed, a total of 4,320 trials per subject per model per comparison (across 15 images and 8 scaling values) for all energy model comparisons and for luminance model original vs. synthesized white noise comparison. Luminance model synthesized vs. synthesized, white noise comparison had 1,440 trials (across 5 images and 8 scaling values) for a single subject. Where behavioral data is plotted in this paper, the proportion correct is the average across all relevant trials.

We fit psychophysical curves describing proportion correct as a function of model scaling using the two-parameter function for discriminability d^t derived in ***Freeman and Simoncelli (2011)***:

$$d'(s;\alpha,s_c)=\begin{cases}\alpha(1-\frac{s_c^2}{s^2}), & s>s_c\\ 0, & s\le s_c\end{cases} \quad (1)$$

where s_c is the critical scaling value (performance is at chance for scaling values at or below s_c) and a is the max d^t value (called the "proportionality factor" in ***Freeman and Simoncelli (2011)***).

Psychophysical curves were constructed by converting this d^t into the probability correct using the same function as in ***Freeman and Simoncelli (2011)***:

$$P(s;\alpha,s_c)=\Phi\left(\frac{d'(s;\alpha,s_c)}{\sqrt{2}}\right)\Phi\left(\frac{d'(s;\alpha,s_c)}{2}\right)+\Phi\left(\frac{-d'(s;\alpha,s_c)}{\sqrt{2}}\right)\Phi\left(\frac{-d'(s;\alpha,s_c)}{2}\right) \quad (2)$$

where φ is the cumulative of the normal distribution. The probability correct is 50% when $d^{'} = 0$ (and thus when scaling is at or below the critical scaling), reaches about 79% when $d^{'} = 2$ and 98% when $d^{'} = 4$. As the a parameter above gives the maximum $d^{'}$ value, it has a monotonic relationship with the asymptotic performance, which can be seen in **figure 16**.

The posterior distribution over parameters s_c and a was estimated using a hierarchical, partial-pooling model, with independent subject- and image-level effects for both s_c and a, with each model and comparison estimated separately, following the procedure used in ***Wallis et al. (2019)***. Subject responses were modeled as samples from a Bernoulli distribution with probability $(1 - 1r)P(s)+.51r$, where $1r$ is the lapse rate, estimated independently for each subject. Estimates were obtained using a Markov Chain Monte Carlo (MCMC) procedure written in Python 3.7.10 (***Van Rossum and Drake, 2009***) using the numpyro package, version 0.8.0 (***Phan et al., 2019***; ***Bingham et al., 2018***). MCMC sampling was conducted using the No U-Turn Sampler algorithm (***Hoffman and Gelman, 2014***), with parameters selected to ensure convergence, which was assessed using the $\hat{R}$ statistics (***Brooks and Gelman (1998)***, looking for $\hat{R} < 1.01 < 1.01$, ***Vehtari et al. (2021)***) and by examining traceplots.

Parameters were given weakly-informative priors and both s_c and a were estimated on natural logarithmic scales.

In sum, for model m E {E, L}, comparison t, subject x, image i, and scaling s:

$$y_1,\ldots,y_n \sim \text{Bernoulli}((1-\pi_{mtx})P(s;\alpha_{mtxi},s_{c,mtxi})+.5\pi_{mtx}) \quad (3)$$

$$\log\alpha_{mtxi}=\alpha_{mt}+\alpha_{mti}+\alpha_{mtx} \quad (4)$$

$$\log s_{c,mtxi}=s_{c,mt}+s_{c,mti}+s_{c,mtx} \quad (5)$$

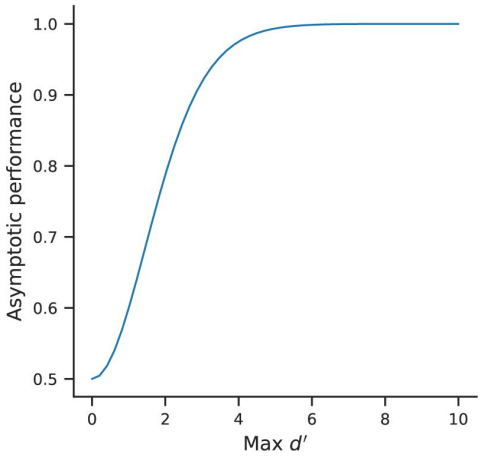


Figure 16.

Relationship between the max *d*t parameter, *a* and asymptotic performance. As max *d*t increases beyond approximately 5 (where asymptotic performance is at ceiling), the slope of the psychophysical curve continues to increase (for example, compare the slope of the luminance and energy model original vs. synth white noise comparisons in **figure 9A**).

with the following priors:

$$\alpha_{mt} \sim \mathcal{N}(1.6, 1)$$
$$s_{c,Et} \sim \mathcal{N}(-1.38, 1)$$
$$s_{c,Lt} \sim \mathcal{N}(-4, 1)$$
$$\pi_{mtx} \sim \text{Beta}(2, 50)$$
$$\alpha_{mtx} \sim \mathcal{N}(0, \sigma_{\alpha,mtx})$$
$$\alpha_{mti} \sim \mathcal{N}(0, \sigma_{\alpha,mti})$$
$$s_{c,mtx} \sim \mathcal{N}(0, \sigma_{s_c,mtx})$$
$$s_{c,mti} \sim \mathcal{N}(0, \sigma_{s_c,mti})$$
$$\sigma_{\alpha,mtx} \sim \text{HalfCauchy}(.1)$$
$$\sigma_{\alpha,mti} \sim \text{HalfCauchy}(.1)$$
$$\sigma_{s_c,mtx} \sim \text{HalfCauchy}(.1)$$
$$\sigma_{s_c,mti} \sim \text{HalfCauchy}(.1)$$

The priors for $s_{c,mt}$ of the energy and luminance models correspond to critical scales of 0.25 and 0.018, respectively, the centers of the V1 physiological range provided in ***Freeman and Simoncelli (2011)*** **figure 5**, and from the slope of a line fit to the dendritic field diameter vs. eccentricity of midget retinal ganglion cells in ***Dacey and Petersen (1992)*** **figure 2B** (see appendix section 2). This captures our prediction that the models' critical scaling values should be similar to those of the physiological scaling in the brain area sensitive to the same image features, should be independent of comparison type and consistent across images and subjects, while not placing too much of a constraint on the parameters.

The posterior distribution represents the model's beliefs about the parameters given the priors and data and is summarized throughout this paper as the posterior mean and 95% high density intervals. The latter represents the range of values containing 95% of the distribution with the highest probability, as opposed to the more common 95% confidence interval, which is symmetrically arranged around the mean. The two are identical for symmetric distributions, but can diverge markedly if the distribution is highly skewed (***Kruschke, 2015***).

Software

These experiments relied on a variety of custom scripts written in Python 3.7.10 (***Van Rossum and Drake, 2009***), all found in the GitHub repository associated with this paper. The following packages were used: snakemake (***Mölder et al., 2021***), JAX (***Bradbury et al., 2018***), matplotlib (***Hunter, 2007***), psychopy (***Peirce et al., 2019***), scipy (***Virtanen et al., 2020***), scikit-image (***van der Walt et al., 2014***), pytorch (***Paszke et al., 2019***), arviz (***Kumar et al., 2019***), numpyro

(*Phan et al., 2019*; *Bingham et al., 2018*), pandas (*Reback et al., 2021*; *McKinney, 2010*), seaborn (*Waskom, 2021*), jupyterlab (*Kluyver et al., 2016*), and xarray (*Hoyer and Hamman, 2017*).

Acknowledgements

The authors would like to thank David Brainard for the use of his photographs, both from the published UPenn Natural Image Database (*Tkačik et al., 2011*) and the unpublished set of images from around Philadelphia. They would also like to thank Tony Movshon, David Heeger, David Brainard, Corey Ziemba, and Colin Bredenberg for their feedback on the manuscript, Mike Landy for his assistance with the design of the psychophysical task and feedback on the manuscript, Heiko Schütt for his assistance with the Markov Chain Monte Carlo analysis, and the authors of *Wallis et al. (2019)* for sharing their code and data. Furthermore, they would like to thank Liz Lovero, Paul Murray, Dylan Simon, and Aaron Watters for their work in creating the metamer browser website.

References

Anderson SJ, Mullen KT, Hess RF (1991) **Human Peripheral Spatial Resolution for Achromatic and Chromatic Stimuli: Limits Imposed By Optical and Retinal Factors** *The Journal of Physiology* **442**:47–64

Anstis SM (1974) **A Chart Demonstrating Variations in Acuity With Retinal Position** *Vision Research* **14**:589–592 https://doi.org/10.1016/0042-6989(74)90049-2

Anstis S. (1998) **Picturing Peripheral Acuity** *Perception* **27**:817–825 https://doi.org/10.1068/p270817

Balas B, Nakano L, Rosenholtz R (2009) **A Summary-Statistic Representation in Peripheral Vision Explains Visual Crowding** *Journal of Vision* **9**:13–13 https://doi.org/10.1167/9.12.13

Banks MS, Geisler WS, Bennett PJ (1987) **The Physical Limits of Grating Visibility** *Vision Research* **27**:1915–1924 https://doi.org/10.1016\%2F0042-6989\%2887\%2990057-5

Bennett PJ, Cortese F (1996) **Masking of Spatial Frequency in Visual Memory Depends on Distal, Not Retinal, Frequency** *Vision Research* **36**:233–238 https://doi.org/10.1016/0042-6989(95)00085-e

Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman ND (2018) **Pyro: Deep Universal Probabilistic Programming** *arXiv preprint arXiv:181009538*

Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S, Zhang Q (2018) **Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S, Zhang Q, JAX: composable transformations of Python+NumPy programs; 2018. http://github.com/google/jax.**

Brooks SP, Gelman A (1998) **General Methods for Monitoring Convergence of Iterative Simulations** *Journal of Computational and Graphical Statistics* **7**:434–455 https://doi.org/10.1080/10618600.1998.10474787

Brown R, DuTell V, Walter B, Rosenholtz R, Shirley P, McGuire M, Luebke D (2021) **Effcient Dataflow Modeling of Peripheral Encoding in the Human Visual System**

Bruna J, Mallat S (2013) **Invariant Scattering Convolution Networks** *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**:1872–1886 https://doi.org/10.1109%2Ftpami.2012.230

Burt PJ, Adelson EH (1987) **The Laplacian pyramid as a compact image code** *Readings in computer vision* :671–679

Cohen JB, Kappauf WE (1985) **Color Mixture and Fundamental Metamers: Theory, Algebra, Geometry, Application** *The American Journal of Psychology* **98** https://doi.org/10.2307/1422442

Dacey DM, Petersen MR (1992) **Dendritic Field Size and Morphology of Midget and Parasol Ganglion Cells of the Human Retina** *Proceedings of the National Academy of Sciences* **89**:9666–9670 https://doi.org/10.1073/pnas.89.20.9666

Daniel PM, Whitteridge D (1961) **The Representation of the Visual Field on the Cerebral Cortex in Monkeys** *The Journal of Physiology* **159**:203–221 https://doi.org/10.1113%2Fjphysiol.1961.sp006803

Deza A, Jonnalagadda A, Eckstein MP (2019) **Deza A, Jonnalagadda A, Eckstein MP. Towards Metamerism via Foveated Style Transfer. In: International Conference on Learning Representations; 2019. https://openreview.net/forum?id=BJzbG20cFQ.**

Douglas RJ, Martin KAC, Whitteridge D (1989) **A Canonical Microcircuit for Neocortex** *Neural Computation* **1**:480–488 https://doi.org/10.1162%2Fneco.1989.1.4.480

Duncan RO, Boynton GM (2003) **Cortical Magnification Within Human Primary Visual Cortex Correlates With Acuity Thresholds** *Neuron* **38**:659–671 https://doi.org/10.1016/s0896-6273(03)00265-4

Feather J, Durango A, Gonzalez R, McDermott J (2019) **Metamers of neural networks reveal divergence from human perceptual systems** *NeurIPS* :10078–10089

Freeman J, Simoncelli EP (2011) **Metamers of the ventral stream** *Nature Neuroscience* **14**:1195–1201 https://doi.org/10.1038/nn.2889

Frisen L, Glansholm A (1975) **Optical and Neural Resolution in Peripheral Vision** *Investigative Ophthalmology & Visual Science* **14**:528–536

Fukushima K (1980) **Neocognitron: a Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected By Shift in Position** *Biological Cybernetics* **36**:193–202 https://doi.org/10.1007%2Fbf00344251

Gattass R, Gross CG, Sandell JH (1981) **Visual Topography of V2 in the Macaque** *The Journal of Comparative Neurology* **201**:519–539 https://doi.org/10.1002/cne.902010405

Gattass R, Sousa A, Gross C (1988) **Visuotopic Organization and Extent of V3 and V4 of the Macaque** *Journal of Neuroscience* **8**:1831–1845

Greenwood JA, Bex PJ, Dakin SC (2009) **Positional Averaging Explains Crowding With Letter-Like Stimuli** *Proceedings of the National Academy of Sciences* **106**:13130–13135 https://doi.org/10.1073%2Fpnas.0901352106

Heeger DJ, Simoncelli EP, Movshon JA (1996) **Computational Models of Cortical Visual Processing** *Proceedings of the National Academy of Sciences* **93**:623–627

Helmholtz H. LXXXI (1852) **On the Theory of Compound Colours** *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **4**:519–534 https://doi.org/10.1080/14786445208647175

Hoffman MD, Gelman A (2014) **The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo** *Journal of Machine Learning Research* **15**:1593–1623

Hoyer S, Hamman J (2017) **Xarray: N-D Labeled Arrays and Datasets in Python** *Journal of Open Research Software* **5** https://doi.org/10.5334/jors.148

Hunter JD (2007) **Matplotlib: a 2d Graphics Environment** *Computing in Science & Engineering* **9**:90–95 https://doi.org/10.1109/MCSE.2007.55

Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG (2018) **Averaging Weights Leads To Wider Optima and Better Generalization** *arXiv preprint arXiv:180305407*

Jagadeesh AV, Gardner JL (2022) **Texture-Like Representation of Objects in Human Visual Cortex** *bioRxiv* https://doi.org/10.1101/2022.01.04.474849

Keshvari S, Rosenholtz R (2016) **Pooling of Continuous Features Provides a Unifying Account of Crowding** *Journal of Vision* **16** https://doi.org/10.1167/16.3.39

Kingma DP, Ba J. (2014) **Adam: A Method for Stochastic Optimization** *ArXiv e-prints*

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, development team J, Loizides F, Scmidt B (2016) **Jupyter Notebooks - a publishing format for reproducible computational workflows** *Positioning and Power in Academic Publishing: Players, Agents and Agendas* :87–90

Kruschke JK (2015) **Doing Bayesian Data Analysis** *Second ed. Elsevier* https://doi.org/10.1016/c2012-0-00477-2

Kumar R, Carroll C, Hartikainen A, Martin O (2019) **Arviz a Unified Library for Exploratory Analysis of Bayesian Models in Python** *Journal of Open Source Software* **4** https://doi.org/10.21105/joss.01143

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) **Backpropagation Applied To Handwritten Zip Code Recognition** *Neural Computation* **1**:541–551 https://doi.org/10.1162/neco.1989.1.4.541

Lettvin JY (1976) **On Seeing Sidelong** *The Sciences* **16**:10–20 https://doi.org/10.1002%2Fj.2326-1951.1976.tb01231.x

Maunsell JHR, Newsome WT (1987) **Visual Processing in Monkey Extrastriate Cortex** *Annual Review of Neuroscience* **10**:363–401 https://doi.org/10.1146%2Fannurev.ne.10.030187.002051

McKinney W, van der Walt S, Millman J (2010) **Data Structures for Statistical Computing in Python** :56–61 https://doi.org/10.25080/Majora-92bf1922-00a

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J (2021) **Sustainable Data Analysis With Snakemake** *F1000Research* **10** https://doi.org/10.12688/f1000research.29032.2

Parkes L, Lund J, Angelucci A, Solomon JA, Morgan M (2001) **Compulsory Averaging of Crowded Orientation Signals in Human Vision** *Nature Neuroscience* **4**:739–744 https://doi.org/10.1038%2F89532

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, et al., Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (2019) **PyTorch: An Imperative Style, High-Performance Deep Learning Library** *Advances in Neural Information Processing Systems* :8024–8035

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK (2019) **PsychoPy2: Experiments in Behavior Made Easy** *Behavior Research Methods* **51**:195–203 https://doi.org/10.3758/s13428-018-01193-y

Pelli DG, Palomares M, Majaj NJ (2004) **Crowding Is Unlike Ordinary Masking: Distinguishing Feature Integration From Detection** *Journal of Vision* **4** https://doi.org/10.1167%2F4.12.12

Phan D, Pradhan N, Jankowiak M (2019) **Composable Effects for Flexible and Accelerated Probabilistic Programming in Numpyro** *arXiv preprint arXiv:191211554*

Portilla J, Simoncelli EP (2000) **A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients** *International journal of computer vision* **40**:49–70

Reback J, jbrockmendel, McKinney W, den Bossche JV, Augspurger T, Cloud P, Hawkins S, gfyoung, Roeschke M, Sinhrks, et al (2021) **Reback J, jbrockmendel, McKinney W, den Bossche JV, Augspurger T, Cloud P, Hawkins S, gfyoung, Roeschke M, Sinhrks, et al, pandas-dev/pandas: Pandas 1.2.3. Zenodo; 2021. doi: 10.5281/zenodo.4572994.** https://doi.org/10.5281/zenodo.4572994

Reddi SJ, Kale S, Kumar S (2018) **On the Convergence of Adam and Beyond**

Riesenhuber M, Poggio T (1999) **Hierarchical Models of Object Recognition in Cortex** *Nature Neuroscience* **2**:1019–1025 https://doi.org/10.1038%2F14819

Robson J, Graham N (1981) **Probability Summation and Regional Variation in Contrast Sensitivity Across the Visual Field** *Vision research* **21**:409–418

Rovamo J, Virsu V, Näsänen R (1978) **Cortical Magnification Factor Predicts the Photopic Contrast Sensitivity of Peripheral Vision** *Nature* **271**:54–56 https://doi.org/10.1038/271054a0

Schnapf JL, Kraft TW, Baylor DA (1987) **Spectral Sensitivity of Human Cone Photoreceptors** *Nature* **325**:439–441 https://doi.org/10.1038%2F325439a0

Schwartz EL (1977) **Spatial Mapping in the Primate Sensory Projection: Analytic Structure and Relevance To Perception** *Biological Cybernetics* **25**:181–194 https://doi.org/10.1007%2Fbf01885636

Simoncelli EP, Freeman WT (1995) **The Steerable Pyramid: A flexible architecture for multi-scale derivative computation** :444–447 https://doi.org/10.1109/ICIP.1995.537667

Song S, Levi DM, Pelli DG (2014) **A Double Dissociation of the Acuity and Crowding Limits To Letter Identification, and the Promise of Improved Visual Screening** *Journal of Vision* **14**:3–3 https://doi.org/10.1167/14.5.3

Teller DY (1984) **Linking Propositions** *Vision research* **24**:1233–1246

Thibos LN (2020) **Retinal Image Formation and Sampling in a Three-Dimensional World** *Annual Review of Vision Science* **6**:469–489 https://doi.org/10.1146/annurev-vision-121219-081840

Tkačik G, Garrigan P, Ratliff C, Milčinski G, Klein JM, Seyfarth LH, Sterling P, Brainard DH, Balasubramanian V (2011) **Natural Images From the Birthplace of the Human Eye** *PLoS ONE* **6** https://doi.org/10.1371%2Fjournal.pone.0020409

Van Rossum G, Drake FL (2009) **Python 3 Reference Manual. Scotts Valley**

Vehtari A, Gelman A, Gabry J (2016) **Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC** *Statistics and Computing* **27**:1413–1432 https://doi.org/10.1007/s11222-016-9696-4

Vehtari A, Gelman A, Simpson D, Carpenter B, Rank-Normalization Bürkner PC. (2021) **and Localization: an Improved R" for Assessing Convergence of Mcmc (with Discussion)** *Bayesian Analysis* **16**:667–718 https://doi.org/10.1214/20-BA1221

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Jarrod Millman K, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey C, et al. (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python** *Nature Methods* **17**:261–272 https://doi.org/https://doi.org/10.1038/s41592-019-0686-2

Wallis TS, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M (2019) **Image Content Is More Important Than Bouma's Law for Scene Metamers** *eLife* **8** https://doi.org/10.7554/elife.42512

van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E (2014) **Yu T, the scikit-image contributors. Scikit-Image: Image Processing in Python** *PeerJ* **6** https://doi.org/10.7717/peerj.453

Wandell BA (1995) **Wandell BA. Foundations of vision. Sinauer Associates; 1995. https://foundationsofvision.stanford.edu/.**

Wandell BA, Winawer J (2015) **Computational Neuroimaging and Population Receptive Fields** *Trends in cognitive sciences* **19**:349–357

Wang Z, Bovik AC (2009) **Mean Squared Error: Love It Or Leave It? A New Look At Signal Fidelity Measures** *IEEE Signal Processing Magazine* **26**:98–117 https://doi.org/10.1109%2Fmsp.2008.930649

Waskom ML (2021) **Seaborn: Statistical Data Visualization** *Journal of Open Source Software* **6** https://doi.org/10.21105/joss.03021

Watanabe S (2013) **A Widely Applicable Bayesian Information Criterion** *Journal of Machine Learning Research* **14**:867–897

Watson AB, Ahumada AJ, Farrell JE (1986) **Window of Visibility: a Psychophysical Theory of Fidelity in Time-Sampled Visual Motion Displays** *JOSA A* **3**:300–307

Ziemba CM, Simoncelli EP (2021) **Opposing Effects of Selectivity and Invariance in Peripheral Vision** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-24880-5

Article and author information

William F. Broderick
Flatiron Institute, Simons Foundation
For correspondence: billbrod@gmail.com
ORCID iD: 0000-0002-8999-9003

Gizem Rufo
Meta, Inc.

Jonathan Winawer
Department of Psychology, New York University
ORCID iD: 0000-0001-7475-5586

Eero P. Simoncelli
Flatiron Institute, Simons Foundation, Center for Neural Science, New York University, Courant Inst. for Mathematical Sciences, New York University
ORCID iD: 0000-0002-1206-527X

Copyright

Editors

Reviewing Editor
Floris de Lange
Donders Institute for Brain, Cognition and Behaviour, Netherlands

Senior Editor
Floris de Lange
Donders Institute for Brain, Cognition and Behaviour, Netherlands

Reviewer #1 (Public Review):

This is an interesting study of the nature of representations across the visual field. The question of how peripheral vision differs from foveal vision is a fascinating and important one. The majority of our visual field is extra-foveal yet our sensory and perceptual capabilities decline in pronounced and well-documented ways away from the fovea. Part of the decline is thought to be due to spatial averaging ('pooling') of features. Here, the authors contrast two models of such feature pooling with human judgments of image content. They use much larger visual stimuli than in most previous studies, and some sophisticated image synthesis methods to tease apart the prediction of the distinct models.

More importantly, in so doing, the researchers thoroughly explore the general approach of probing visual representations through metamers-stimuli that are physically distinct but perceptually indistinguishable. The work is embedded within a rigorous and general mathematical framework for expressing equivalence classes of images and how visual representations influence these. They describe how image-computable models can be used to make predictions about metamers, which can then be compared to make inferences about the underlying sensory representations. The main merit of the work lies in providing a formal framework for reasoning about metamers and their implications, for comparing models of sensory processing in terms of the metamers that they predict, and for mapping such models onto physiology. Importantly, they also consider the limits of what can be inferred about sensory processing from metamers derived from different models.

Overall, the work is of a very high standard and represents a significant advance over our current understanding of perceptual representations of image structure at different locations across the visual field. The authors do a good job of capturing the limits of their approach and I particularly appreciated the detailed and thoughtful Discussion section and the suggestion to extend the metamer-based approach described in the MS with observer models. The work will have an impact on researchers studying many different aspects of visual function including texture perception, crowding, natural image statistics, and the physiology of low- and mid-level vision.

The main weaknesses of the original submission relate to the writing. A clearer motivation could have been provided for the specific models that they consider, and the text could have been written in a more didactic and easy-to-follow manner. The authors could also have been more explicit about the assumptions that they make.

Reviewer #2 (Public Review):

Summary
This paper expands on the literature on spatial metamers, evaluating different aspects of spatial metamers including the effect of different models and initialization conditions, as well as the relationship between metamers of the human visual system and metamers for a model. The authors conduct psychophysics experiments testing variations of metamer synthesis parameters including type of target image, scaling factor, and initialization parameters, and also compare two different metamer models (luminance vs energy). An additional contribution is doing this for a field of view larger than has been explored previously.

General Comments
Overall, this paper addresses some important outstanding questions regarding comparing original to synthesized images in metamer experiments and begins to explore the effect of noise vs image seed on the resulting syntheses. While the paper tests some model classes that could be better motivated, and the results are not particularly groundbreaking, the contributions are convincing and undoubtedly important to the field. The paper includes an interesting Voronoi-like schematic of how to think about perceptual metamers, which I found helpful, but for which I do have some questions and suggestions. I also have some major concerns regarding incomplete psychophysical methodology including lack of eye-tracking, results inferred from a single subject, and a huge number of trials. I have only minor typographical criticisms and suggestions to improve clarity. The authors also use very good data reproducibility practices.

Specific Comments

Experimental Setup
Firstly, the experiments do not appear to utilize an eye tracker to monitor fixation. Without eye tracking or another manipulation to ensure fixation, we cannot ensure the subjects were fixating the center of the image, and viewing the metamer as intended. While the short stimulus time (200ms) can help minimize eye movements, this does not guarantee that subjects began the trial with correct fixation, especially in such a long experiment. While Covid-19 did at one point limit in-person eye-tracked experiments, the paper reports no such restrictions that would have made the addition of eye-tracking impossible. While such a large-scale experiment may be difficult to repeat with the addition of eye tracking, the paper would be greatly improved with, at a minimum, an explanation as to why eye tracking was not included.

Secondly, many of the comparisons later in the paper (Figures 9,10) are made from a single subject. N=1 is not typically accepted as sufficient to draw conclusions in such a

psychophysics experiment. Again, if there were restrictions limiting this it should be discussed. Also (P11) Is subject sub-00 is this an author? Other expert? A naive subject? The subject's expertise in viewing metamers will likely affect their performance.

Finally, the number of trials per subject is quite large. 13,000 over 9 sessions is much larger than most human experiments in this area. The reason for this should be justified.

Model
For the main experiment, the authors compare the results of two models: a 'luminance model' that spatially pools mean luminance values, and an 'energy model' that spatially pools energy calculated from a multi-scale pyramid decomposition. They show that these models create metamers that result in different thresholds for human performance, and therefore different critical scaling parameters, with the basic luminance pooling model producing a scaling factor 1/4 that of the energy model. While this is certain to be true, due to the luminance model being so much simpler, the motivation for the simple luminance-based model as a comparison is unclear.

The authors claim that this luminance model captures the response of retinal ganglion cells, often modeled as a center-surround operation (Rodieck, 1964). I am unclear in what aspect(s) the authors claim these center-surround neurons mimic a simple mean luminance, especially in the context of evidence supporting a much more complex role of RGCs in vision (Atick & Redlich, 1992). Why do the authors not compare the energy model to a model that captures center-surround responses instead? Do the authors mean to claim that the luminance model captures only the pooling aspects of an RGC model? This is particularly confusing as Figures 6 and 9 show the luminance and energy models for original vs synth aligning with the scaling of Midget and Parasol RGCs, respectively. These claims should be more clearly stated, and citations included to motivate this. Similarly, with the energy model, the physiological evidence is very loosely connected to the model discussed.

Prior Work:
While the explorations in this paper clearly have value, it does not present any particularly groundbreaking results, and those reported are consistent with previous literature. The explorations around critical eccentricity measurement have been done for texture models (Figure 11) in multiple papers (Freeman 2011, Wallis, 2019, Balas 2009). In particular, Freeman 20111 demonstrated that simpler models, representing measurements presumed to occur earlier in visual processing need smaller pooling regions to achieve metamerism. This work's measurements for the simpler models tested here are consistent with those results, though the model details are different. In addition, Brown, 2023 (which is miscited) also used an extended field of view (though not as large as in this work). Both Brown 2023, and Wallis 2019 performed an exploration of the effect of the target image. Also, much of the more recent previous work uses color images, while the author's exploration is only done for greyscale.

Discussion of Prior Work:
The prior work on testing metamerism between original vs. synthesized and synthesized vs. synthesized images is presented in a misleading way. Wallis et al.'s prior work on this should not be a minor remark in the post-experiment discussion. Rather, it was surely a motivation for the experiment. The text should make this clear; a discussion of Wallis et al. should appear at the start of that section. The authors similarly cite much of the most relevant literature in this area as a minor remark at the end of the introduction (P3L72).

White Noise:
The authors make an analogy to the inability of humans to distinguish samples of white noise. It is unclear however that human difficulty distinguishing samples of white noise is a perceptual issue- It could instead perhaps be due to cognitive/memory limitations. If one concentrates on an individual patch one can usually tell apart two samples. Support for these

difficulties emerging from perceptual limitations, or a discussion of the possibility of these limitations being more cognitive should be discussed, or a different analogy employed.

Relatedly, in Figure 14, the authors do not explain why the white noise seeds would be more likely to produce syntheses that end up in different human equivalence classes.

It would be nice to see the effect of pink noise seeds, which mirror the power spectrum of natural images, but do not contain the same structure as natural images - this may address the artifacts noted in Figure 9b.

Finally, the authors note high-frequency artifacts in Figure 4 & P5L135, that remain after syntheses from the luminance model. They hypothesize that this is due to a lack of constraints on frequencies above that defined by the pooling region size. Could these be addressed with a white noise image seed that is pre-blurred with a low pass filter removing the frequencies above the spatial frequency constrained at the given eccentricity?

Schematic of metamerism:
Figures 1,2,12, and 13 show a visual schematic of the state space of images, and their relationship to both model and human metamers. This is depicted as a Voronoi diagram, with individual images near the center of each shape, and other images that fall at different locations within the same cell producing the same human visual system response. I felt this conceptualization was helpful. However, implicitly it seems to make a distinction between metamerism and JND (just noticeable difference). I felt this would be better made explicit. In the case of JND, neighboring points, despite having different visual system responses, might not be distinguishable to a human observer.

In these diagrams and throughout the paper, the phrase 'visual stimulus' rather than 'image' would improve clarity, because the location of the stimulus in relation to the fovea matters whereas the image can be interpreted as the pixels displayed on the computer.

Other
The authors show good reproducibility practices with links to relevant code, datasets, and figures.